A Crowd-AI Collaborative Duo Relational Graph Learning Framework towards Social Impact Aware Photo Classification

Yang Zhang[1], Ziyi Kou[2], Lanyu Shang[1], Huimin Zeng[1], Zhenrui Yue[1], Dong Wang[1]

[1] School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL, USA
[2] Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA
yzhangnd@illinois.edu, zkou@nd.edu, {lshang3, huiminz3, zhenrui3, dwang24}@illinois.edu

Abstract

In artificial intelligence (AI), negative social impact (NSI) represents the negative effect on the society as a result of mistakes conducted by AI agents. While the photo classification problem has been widely studied in the AI community, the NSI made by photo misclassification is largely ignored due to the lack of quantitative measurements of the NSI and effective approaches to reduce it. In this paper, we focus on an NSI-aware photo classification problem where the goal is to develop a novel crowd-AI collaborative learning framework that leverages online crowd workers to quantitatively estimate and effectively reduce the NSI of misclassified photos. Our problem is motivated by the limitations of current NSI-aware photo classification approaches that either 1) cannot accurately estimate NSI because they simply model NSI as the semantic difference between true and misclassified categories or 2) require costly human annotations to estimate NSI of pairwise class categories. To address such limitations, we develop SocialCrowd, a crowdsourcing-based NSI-aware photo classification framework that explicitly reduces the NSI of photo misclassification by designing a duo relational NSI-aware graph with the NSI estimated by online crowd workers. The evaluation results on two large-scale image datasets show that SocialCrowd not only reduces the NSI of photo misclassification but also improves the classification accuracy on both datasets.

Introduction

Photo classification is a prevalent image classification application that classifies online photos into semantic categories (Hu et al. 2018). The classified photos are leveraged in various downstream online applications, such as hashtag recommendation for photo sharing services (Zhang et al. 2019), automatic photo organization (Lonn, Radeva, and Dimiccoli 2019), and keyword-based social media image retrieval (Chen et al. 2022). However, such applications often suffer from the misclassification issue where the photos are classified into incorrect categories, and such misclassifications cause severe *negative social impact (NSI)* on the society. For example, an AI model developed by Facebook made a high NSI (e.g., strong public criticism) because it wrongly classified a photo of a black man in a video as content about "Primates" and recommended the photo to users as an advertisement (Wehrli et al. 2021). Formally, we define NSI as negative or undesirable effects on society as a result of misclassifying a given photo by an AI-based photo classification model (e.g., deep convolutional network). Such NSI usually contradicts the mainstream society value or humanity (Chen and Bu 2019). This paper focuses on an NSI-aware photo classification problem where the goal is to design a novel crowd-AI collaborative photo classification framework that effectively minimizes NSI of misclassified photos.

Several initial efforts have been made to study the NSI-related problem caused by photo misclassification in AI and computer vision communities (Sengupta et al. 2018; Olmo, Sengupta, and Kambhampati 2020; Sengupta 2020). In particular, Sengupta *et al.* modeled the negative impact of misclassified photos as the semantic differences between the true photo labels and misclassified labels (Sengupta et al. 2018). For example, they expect an autonomous car to pose a lower negative impact if it misclassifies a photo of a "Dog" as a "Cat" than as a "Plastic Bag". The reason is that the "Dog" and "Cat" sub-categories both belong to the "Animal" category that requires the vehicle to stop while the "Plastic Bag" does not. However, the above semantic relations are insufficient to estimate the NSI of misclassified online photos because such semantic relations are generated without explicit considerations of social ethics and values (Friedman and Kahn Jr 2007). For example, online users feel more offensive if human-related photos are misclassified as specific animals (e.g., "Gorillas", "Pigs") than objects (e.g., "Desk", "Vase") even if the labels of animals share more semantic relation with "Human" as they are all living creatures.

Figure 1 further demonstrates our NSI-aware photo classification problem. We show four misclassified photos with their true and misclassified photo labels in Figure 1. From the perspective of online users who have seen these misclassified photos, we analyze the potential NSI made by each misclassified photo and observe that the misclassification results can lead to different levels of NSI. For example, misclassifying a "Person" as a "Giraffe" (Figure 1a) could make a high NSI because many people feel offended by misclassifying their personal photos as specific animals. In contrast, the NSI of misclassifying a "Poodle Dog" as a "Rabbit"

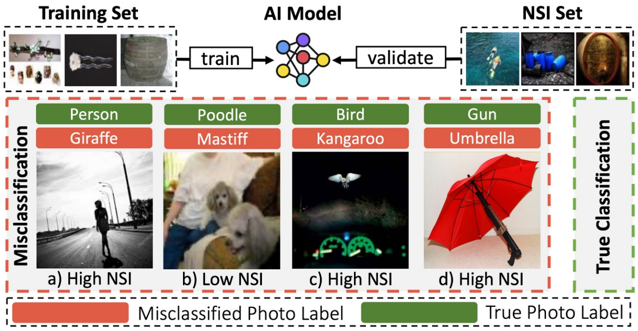


Figure 1: NSI-aware Photo Classification Problem

(Figure 1b) is low since both "Poodle Dog" and "Rabbit" are small animals. However, a close semantic relation between two class categories (e.g., "Poodle Dog" and "Rabbit") does not always represent a low NSI of the misclassified photos between these categories. For example, misclassifying a bird as a kangaroo (Figure 1c) could cause a high NSI because the risk of a car crashing with a kangaroo is significantly higher than a flying bird. Moreover, we also observe that the reason for misclassifying a photo could be closely associated with the ambiguous appearance of objects in the photo. For example, Figure 1d contains a real gun that looks like an umbrella as the gun is hidden by installing an umbrella top frame. Therefore, it is likely that a trained photo classification model misclassifies the gun as "Umbrella" in the photo, leading to a high NSI due to public safety concerns.

Motivated by the above observations, we develop SocialCrowd, a crowd-AI collaborative learning framework that leverages the collaborative strengths of AI and crowdsourced human intelligence to address the NSI-aware photo classification problem. Our goal is to accurately estimate the NSI observed in misclassified photos from the NSI set and minimize the possibility of classifying unseen photos to the wrong class categories with high NSI. In our SocialCrowd framework, we leverage human intelligence from crowdsourcing systems (e.g., Amazon MTurk) to quantitatively estimate the NSI of misclassified photos. Our motivation for incorporating crowdsourcing efforts is to leverage the crowd workers' extensive background knowledge and experiences to accurately estimate NSI based on their considerations of the relevant social contexts (Zuccon et al. 2011; Savenkov, Weitzner, and Agichtein 2016; Zhang et al. 2021b). The estimated NSI values from different class categories are then leveraged to construct an NSI-aware graph network to accurately classify photos with minimized misclassification NSI. To our best knowledge, SocialCrowd is the first NSI-aware photo classification framework that leverages crowdsourcing intelligence to effectively minimize NSI of misclassified photos across different class categories in large-scale photo classification applications. We evaluate SocialCrowd on two large-scale image classification datasets that contain images from hundreds of different class categories. The results on both datasets show that SocialCrowd significantly reduces the NSI of misclassified photos while accurately classifying the photos compared to state-of-the-art baseline models.

Related Work

Social Impact of AI

The negative social impact (NSI) introduced by AI models remains a challenging problem to be addressed. A few recent efforts have been made to address the NSI issue in AI applications (Wang and Deng 2019; Hu et al. 2021; Korayem et al. 2016). For example, Wang *et al.* developed a skewness-aware reinforcement learning framework to mitigate the racial bias in face recognition applications (Wang and Deng 2019). Korayem *et al.* proposed a privacy-aware object detection framework to reduce the NSI caused by private information on computer screens in photos (Korayem et al. 2016). However, none of the above NSI-aware approaches focuses on addressing the NSI issue caused by photo misclassification due to the lack of effective strategies to estimate and minimize NSI of misclassified photos. There are also recent works that model NSI as semantic differences between class categories of images and develop weighted classification loss functions to reduce NSI (Sengupta et al. 2018; Olmo, Sengupta, and Kambhampati 2020). However, such semantic differences cannot accurately represent NSI and even make wrong estimations on NSI in many real-world scenarios (e.g., the misclassified photo in Figure 1c). Moreover, the above NSI-aware approaches suffer a significant degradation of photo classification performance because they ignore the category ambiguity in NSI-aware photo classification models. In this paper, we focus on a novel NSI-aware large-scale photo classification problem that aims to accurately classify photos and minimize the NSI of misclassified photos by AI models.

Crowd-AI Hybrid Systems

Our work is closely related to crowd-AI hybrid systems where human efforts from a vast amount of crowd workers are collaboratively coordinated to improve the performance of AI algorithms (Hui and Berberich 2017; Blanco et al. 2011; Inel et al. 2018; Shi et al. 2020; Kou et al. 2022b; Zhang et al. 2020, 2022b). For example, Balayn *et al.* introduced a crowdsourcing-based concept extraction approach to interpret image classification results (Balayn et al. 2021). Saralioglu *et al.* developed a post-classification accuracy assessment scheme that leveraged crowdsourcing efforts to evaluate the image classification performance for high-resolution satellite images (Saralioglu and Gungor 2019). Hettiachchi *et al.* designed a visible gold question mechanism to assess the reliability of crowd workers and improve the data quality of crowdsourced face annotations (Hettiachchi et al. 2021). Heim *et al.* proposed a hybrid medical image annotation pipeline that utilized the joint power of the crowd and AI algorithm for medical image segmentation (Heim et al. 2018). However, those solutions do not explore the opportunity to leverage the collective power of crowd and AI to address the NSI-aware photo classification problem. In contrast, the SocialCrowd designs a graph-based crowdsourcing framework that incorporates human intelligence from online crowd workers to accurately estimate the NSI of photo misclassification across different class categories and effectively classify photos with minimal NSI.

Problem Description

We first introduce a few key terms in the crowdsourcing-based NSI-aware photo classification problem.

Definition 1 Photo (x): A photo is an online image that is created and shared by social media users across various social media platforms. The photos containing objects with similar visual characteristics are often classified into the same category. For example, all Dalmatian dogs of different ages and sizes belong to the "Dalmatian" category due to the same breed. We define the categories of interests as $\mathcal{C} = \{c_1, \dots, c_K\}$ with K different categories that are independent of each other. Note that following standard practice in photo classification (Perera, Oza, and Patel 2021), each photo is only assigned to one category based on the main object in that photo.

Definition 2 Photo Dataset ($\mathcal{X}$): A photo dataset contains a set of labeled photos. As shown in Figure 1, the photo dataset $\mathcal{X}$ contains: 1) a **training set** $\mathcal{X}^T = \{x_1^T, \dots, x_M^T\}$ that includes a total of M photos to be used for training AI-related photo classification models; and 2) an **NSI set** $\mathcal{X}^S = \{x_1^S, \dots, x_N^S\}$ that includes a total of N photos to be used for identifying the misclassified photos and evaluating the corresponding NSI values.

Definition 3 Photo Label (y): Each photo x has a semantic photo label $y \in \mathcal{C}$ that identifies the class category of the subjects in x (e.g., $y =$"Person" for the photo in Figure 1a). Similarly, we define the labels of $\mathcal{X}^T$ as $\mathcal{Y}^T = \{y_1^T, \dots, y_M^T\}$ and the labels of $\mathcal{X}^S$ as $\mathcal{Y}^S = \{y_1^S, \dots, y_N^S\}$.

Definition 4 Negative Social Impact (NSI): The negative social impact (NSI) of a photo x represents the negative effect on society if x is misclassified by the AI model (i.e., $\hat{y} \neq y$, where $\hat{y}$ and y are the *predicted* and *ground-truth* photo labels of the image, respectively). In particular, we consider the effect of a misclassified photo as negative if the misclassification contradicts the mainstream society value, common sense, or humanity (Chen and Bu 2019). We denote NSI$(x, y, \hat{y})$ as the quantitative degree of NSI for photo x with the ground truth label y, which is misclassified as $\hat{y}$.

Definition 5 Category Ambiguity (CAB): The category ambiguity of photo x represents the ambiguity degree between class categories of the true and misclassified labels in terms of their visual similarity. For example, the CAB of Figure 1d is high because the gun in the photo is similar to an umbrella and the ambiguity between them is high. We denote $A(x, y, \hat{y})$ as the CAB for x where y and $\hat{y}$ represent the ground-truth and the misclassified labels, respectively.

Definition 6 Duo Relations: Given a misclassified photo, we define the *duo relations* as the NSI and CAB of the photo.

Definition 7 Crowdsourcing Platform ($\mathbb{C}$): A crowdsourcing platform receives crowdsourcing tasks from requesters (e.g., applications) and dispatches the tasks to crowd workers (Zhang et al. 2022a; Kou et al. 2022a). We will illustrate the details of our crowdsourcing task design in the solution.

The goal of our NSI-aware photo classification problem is to accurately classify the photos in $\mathcal{X}$ and minimize the NSI of misclassified photos. Using the definitions above, our problem is formally defined as:

$$
\begin{aligned}
&\text{minimize} \quad \sum_{i=1}^{M} \text{NSI}(x_i, y_i, \hat{y}_i | \text{NSI}(\mathcal{X}^S)), \forall y_i \neq \hat{y}_i \\
&\text{maximize} \quad \sum_{i=1}^{M} \Pr(\hat{y}_i = y_i | \Theta, x_i)
\end{aligned}
\tag{1}
$$

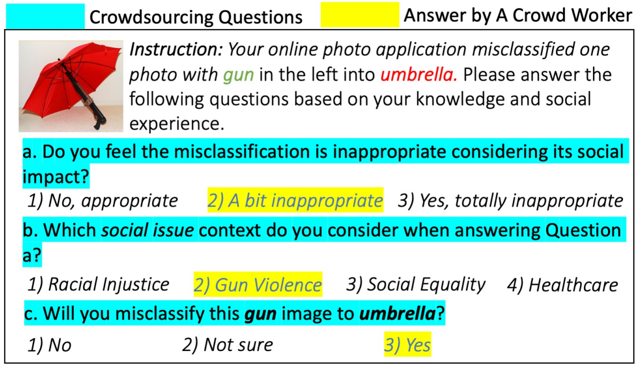
where NSI$(\mathcal{X}^S) = \{\text{NSI}(x_i, y_i, \hat{y}_i) | \mathbb{C}\}, 1 \leq i \leq N$ represents the estimated NSI of the images from the NSI set by the crowdsourcing tasks. $\Pr(\hat{y}_i = y_i | \Theta, x_i)$ is the probability of the AI model Θ to correctly classify photo x_i.

Solution

SocialCrowd consists of three modules: 1) a Crowdsourcing Duo Relational Estimator (CDRE), 2) a Context-driven Visual Relation Predictor (CVRP), and 3) a Graph-based NSI-CAB-aware Classifier (GNCC). In particular, the CDRE module first designs a novel crowdsourcing framework to explicitly query crowdsourced human intelligence to jointly estimate the NSI and CAB of the misclassified photos. The CVRP module, which works in parallel with the CDRE module, designs a metric-based learning-to-learn classification framework to effectively predict the NSI and CAB of the misclassified photos from the categories that are not included in the NSI set. Finally, the GNCC module introduces a graph-based NSI-CAB-aware classifier that carefully fuses the crowdsourced and predicted NSI and CAB from the CDRE and CVRP modules to make accurate photo classification while minimizing the NSI of misclassified photos.

Crowdsourcing Duo Relational Estimator

The crowdsourcing duo relational estimator (CDRE) module aims to leverage crowdsourced human intelligence to estimate both NSI and CAB of misclassified photos across different class categories. CDRE firstly designs a deep misclassified photo identifier that is trained on the training set $\mathcal{X}^T$ and validated on the NSI set $\mathcal{X}^S$. The validation on $\mathcal{X}^S$ identifies the misclassified photos from $\mathcal{X}^S$ by validating the consistency between the predicted and true photo labels. In particular, the identifier trains a set of deep learning photo classification models $\mathcal{M} = \{M_1, \dots, M_F\}$ with different backbone model structures on the different subsets of $\mathcal{X}^T$ and $\mathcal{Y}^T$ based on the cross-entropy loss (Zhang and Sabuncu 2018). The identifier then validates each model in $\mathcal{M}$ on $\mathcal{X}^S$ and $\mathcal{Y}^S$ to identify the misclassified photos. The reason for leveraging a set of photo classification models is that different models with various optimized parameters often generate different classification results on each photo. Therefore, a photo is likely to be misclassified with different labels, which encourages crowd workers to estimate the NSI and CAB of a misclassified photo from different perspectives in the following process. Formally, we define the set of misclassified photos as $\mathcal{X}_*^S = \{x_1^S, \dots, x_{N^*}^S\}$ where $M_f(x_n^S) \neq y_n^S, 1 \leq n \leq N^* \leq N, \forall M_f \in \mathcal{M}$.



🟦 Crowdsourcing Questions 🟨 Answer by A Crowd Worker

Instruction: Your online photo application misclassified one photo with gun in the left into *umbrella*. Please answer the following questions based on your knowledge and social experience.

a. Do you feel the misclassification is inappropriate considering its social impact?

1) No, appropriate *2) A bit inappropriate* *3) Yes, totally inappropriate*

b. Which *social issue* context do you consider when answering Question a?

1) Racial Injustice *2) Gun Violence* *3) Social Equality* *4) Healthcare*

c. Will you misclassify this ***gun*** image to ***umbrella***?

1) No *2) Not sure* *3) Yes*

Figure 2: Visual-guided Duo Relational Estimation Interface

Given the misclassified photos $\mathcal{X}_*^S$ from $\mathcal{X}^S$, CDRE then develops a context-based duo relational crowdsourcing task that is assigned to crowd workers to estimate the NSI and CAB of the photos from $\mathcal{X}_*^S$. In particular, we design a novel visual-guided duo relational estimation interface to interact with crowd workers as shown in Figure 2. For each crowdsourcing task, the interface randomly selects a misclassified photo from $\mathcal{X}_*^S$ and expects a crowd worker to answer three questions by selecting the corresponding options under the questions. In particular, the first and third questions in the interface estimate the NSI and CAB of misclassifying the given photo, respectively. For the second question, we ask crowd workers for the specific context of social issues (i.e., social context) (Borras Jr et al. 2018) that they consider when answering the first question. The motivation of the second question is that crowd workers may estimate the NSI of the misclassified photos from different perspectives based on their own social experiences and interpretations of NSI. Hence, understanding *why* a crowd worker answers the first question with a specific selection is important for aggregating the NSI from different photos and further predicting the NSI of unseen photos in the next subsection. In particular, for each task t with the misclassified photo x_t^S, we define $\{r_{t,i} | 1 \leq i \leq 3\}$ as the set of answers to the three questions.

Context-driven Visual Relation Predictor

As illustrated in Figure 1, the NSI set usually does not contain the misclassified photos with all possible combinations of class categories. We define the missing pairs of class categories as *unobserved category pairs*. In particular, if there is no misclassified photo from the NSI set $\mathcal{X}_S$ that contains (c_i, c_j) as a category pair, we denote (c_i, c_j) as an *unobserved category pair*. For example, if none of the "Kangaroo" photos in the NSI set is misclassified as "Person", ("Kangaroo", "Person") is an unobserved category pair. Since there is no misclassified photo from unobserved category pairs, we cannot estimate the duo relations of the photo by leveraging the CDRE module.

To address the above issue, we develop a metric-based learning-to-learn classification framework to predict the duo relations of misclassified photos from unobserved categories. We show the detailed structure of the framework in Figure 3. To predict the NSI of x_t^S, we firstly generate an embedding matrix $W_E \in \mathbb{R}^{K \times d}$ to transform both c_i and c_j to high-dimensional semantic features $\widetilde{c}_i \in \mathbb{R}^{1 \times d}$ and $\widetilde{c}_j \in \mathbb{R}^{1 \times d}$ where K is the total number of considered class categories. To encode the social context $r_{t,2}$ from R_t as the crowd perspective of NSI, we design a learning-to-learn neural network that transforms the semantic social context into learnable matrix parameters of the framework. The process is denoted as $o_t = W_2(\sigma(W_1(\text{onehot}(r_{t,2}))))$ where $o_t \in \mathbb{R}^{d \times d}$ is the generated parameter matrix. To estimate the potential NSI between c_i and c_j based on the social context $r_{t,2}$, we apply the matrix multiplication for the corresponding features to effectively aggregate the semantic information from the categories c_i, c_j and the social context. The generated features are further encoded with $\widetilde{x}_t^S$ to predict the NSI of x_t^S, where $\widetilde{x}_t^S$ is the encoded high-dimensional vector of x_t^S. The process is denoted as $\hat{r}_{t,1} = \widetilde{x}_t^S \sigma(\widetilde{c}_i^T \widetilde{c}_j o_t) W_3$ where $W_3 \in \mathbb{R}^{d \times 3}$ are learnable parameters to transform the input features to $\hat{r}_{t,1} \in \mathbb{R}^3$ that is optimized with the true NSI value $r_{t,1}$ through cross-entropy loss $\mathcal{L}_{\text{NSI}}(\hat{r}_{t,1}, r_{t,1})$.

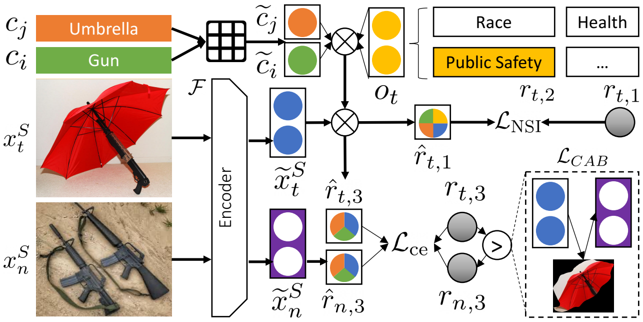


Figure 3: The overview structure of CVRP

To solve the challenge that the CAB of a misclassified photo is associated with the complex visual content of the photo, we design a metric-based pairwise visual feature extractor to effectively identify the representative visual information of different photos. In particular, we first design a deep CAB classifier that classifies the CAB value of $\widetilde{x}_t^S$, which is denoted as $\hat{r}_{t,3} = W_2(\sigma(\{\widetilde{x}_t^S | \widetilde{c}_i | \widetilde{c}_j\} W_1))$ where $\hat{r}_{t,3} \in \mathbb{R}^3$ is the predicted CAB value. We denote the true CAB value $r_{t,3}$ as ground-truth label and apply the cross-entropy loss denoted as $\mathcal{L}_{\text{ce}}(\hat{r}_{t,3}, r_{t,3})$. To further identify the representative visual information related to the corresponding CAB value from each photo embedding, we exchange the visual features between the photo embeddings and expect that the exchanged features lead to the changes of the corresponding CAB values of the photo. In particular, given two photo embeddings $\widetilde{x}_t^S$ and $\widetilde{x}_n^S$ that contain the same category pair (c_i, c_j) but different CAB value $r_{t,3}$ and $r_{n,3}$ where $r_{t,3} > r_{n,3}$, we generate the *exchange visual feature* from $\widetilde{x}_t^S$ as $h_t^S = \max_2(\text{MultiHead}(\widetilde{x}_t^S W_3))$ where MultiHead denotes the multi-head split operation and $\max_2$ denotes the max value extraction on the feature dimension. We then integrate h_t^S with $\widetilde{x}_n^S$ as $H_n^S = \{\{\widetilde{x}_n^S | \widetilde{c}_i | \widetilde{c}_j\} W_1 | h_t^S\}$. Since h_t^S contains specific visual content (e.g., umbrella frame in Figure 2) of photo x_t^S that leads to a high CAB value $r_{t,3}$, the CAB value of H_n^S is expected to be higher than the CAB value (i.e., $r_{n,3}$) from h_n^S because H_n^S contains h_t^S as part of the visual feature. Therefore, we denote the new cross-entropy loss function as $\mathcal{L}_{CAB}(\hat{r}^*_{n,3}, r_{t,3})$, where $\hat{r}^*_{n,3} = W_2 H_n^S$.

After the optimization process of the metric-based learning-to-learn framework based on the loss functions $\mathcal{L}_{\text{NSI}}$, $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{CAB}}$, we leverage the optimized framework to predict the duo relations of each photo from $\mathcal{X}^S$. In particular, we denote the predicted NSI and CAB values of the NSI set $\mathcal{X}^S$ as $\Phi = \{\phi_1, \ldots, \phi_N\}$ and $\Omega = \{\omega_1, \ldots, \omega_N\}$ where N is the total number of photos. $\phi_n = \{\phi_{n,1}, \ldots, \phi_{n,K}\}$ and $\omega_n = \{\omega_{n,1}, \ldots, \omega_{n,K}\}$ denote the predicted NSI and CAB of the photo x_t^S that is misclassified to the total K categories.

Graph-based NSI-CAB-Aware Classifier

We develop the graph-based NSI-CAB-aware classifier (GNCC) to classify the input photos based on the crowd-sourced duo relations of $\mathcal{X}^S$ from the CDRE and the predicted duo relations of $\mathcal{X}^S$ from the CVRP, respectively.

In particular, we first construct a duo relational directed graph (DRDG) as $\mathbb{G}$ based on the duo relations of $\mathcal{X}^S$. In particular, $\mathbb{G}$ contains K graph nodes $\mathcal{V} = \{v_1, \ldots, v_K\}$ that correspond to K class categories. To represent each graph node from $\mathbb{G}$, we design two types of graph embeddings by exploring the semantic meaning of each class category and visual characteristics of the photos with the class category as photo labels. To generate the semantic embeddings of the graph nodes, we create an embedding matrix that transforms each graph node in $\mathcal{V}$ to d-dimensional embeddings. Therefore, the semantic embeddings of all graph nodes can be denoted as $\widetilde{\mathcal{V}}_P = \{\widetilde{v}_{p,1}, \ldots, \widetilde{v}_{p,K}\} \in \mathbb{R}^{K \times d}$. To generate the visual embeddings of the graph nodes, we first encode all photos from $\mathcal{X}^S$ by applying the photo embedding module $\mathcal{F}$ from CVRP and then aggregate the photo embeddings with the same photo labels to generate the embedding for each category. We formally define the visual embeddings of all graph nodes as $\widetilde{\mathcal{V}}_A = \{\widetilde{v}_{a,1}, \ldots, \widetilde{v}_{a,K}\} \in \mathbb{R}^{K \times d}$. Therefore, we denote the node embedding of all graph nodes as $\widetilde{\mathcal{V}} = \{\widetilde{v}_1, \ldots, \widetilde{v}_K\}$ where $\widetilde{v}_k = \{\widetilde{v}_{p,k} | \widetilde{v}_{a,k}\}$ denotes the graph embedding of the graph node v_k. For each pair of nodes (i.e., class categories) v_i and v_j from $\mathbb{G}$, we add a directed graph edge from v_i to v_j to indicate misclassified photos from $\mathcal{X}^S$ with the true label as v_i and the misclassified label as v_j. For each graph edge from v_i to v_j, we define the *NSI edge weight* as the average NSI value of all photos from Φ if the photos have v_i as the photo label and v_j as the misclassified label. Similarly, we define the *CAB edge weight* as the average CAB value of all photos from Ω. We convert the values of all NSI edge weights and CAB edge weights in $\mathbb{G}$ to binary values (i.e., 0 and 1) that represent low and high values of duo relations between different categories based on hyper-thresholds.

Given the constructed $\mathbb{G}$ and an input photo for classification, GNCC aims to aggregate both the photo embedding and graph embeddings based on the structure of $\mathbb{G}$ to discriminate critical information of each class category from $\mathbb{G}$. For example, an input photo with the "Gun" label is more likely to be correctly classified if the classification model can discriminate critical visual information in the photo by explicitly exploring the potential category information from $\mathbb{G}$ that includes the "Gun" category and its connected misclassified categories. Given the input photo $x_m \in \mathcal{X}^T$, we define the graph aggregation process below.

$$\widetilde{h}_k^{(l)} = \{\widetilde{x}_m | \sigma(W_1 \widetilde{v}_k^{(l-1)} + \sum_{j \in \mathcal{V}_k} \alpha_{k,j} \widetilde{v}_j^{(l-1)} W_2)\} \tag{2}$$

where $\widetilde{v}_k^{(l-1)}$ is the aggregated embedding for the k^{th} graph node in $(l-1)^{th}$ graph layer. $\mathcal{V}_k$ denotes the set of graph nodes that are connected with v_k in either or both directions. $\alpha_{k,j}$ is the normalized attention score $A = W_4(\widetilde{\mathcal{V}}(\widetilde{x}_m W_3)^T + (\widetilde{x}_m W_3)\widetilde{\mathcal{V}}^T)$ between k^{th} and j^{th} graph node embeddings. We aggregate the graph node embeddings with the input photo embedding for ζ times where ζ is the pre-defined hyper-parameter.

Given the aggregated embedding $\widetilde{h}_k$ of the graph node v_k from $\mathbb{G}$, our strategy expects $\widetilde{h}_k$ to have larger or smaller distance with other graph node embeddings that are connected to v_k with a higher or lower edge-level NSI, respectively. The strategy increases the probability of the model to consider the categories with low NSI as alternative classification results for the category of v_c. Similarly, our strategy minimizes the distance between $\widetilde{h}_k$ and other graph node embeddings that are connected by graph edges with a higher CAB. For example, our strategy encourages GNCC to jointly consider the graph node embeddings of the "Gun" and "Umbrella" categories to discriminate the ambiguous visual information between the two categories. We define the process of GNCC as follows.

$$\begin{aligned}
\mathcal{L}_p &= \max(\sum_{\alpha \in \mathcal{V}_p^-} \|\widetilde{h}_k - \widetilde{h}_\alpha\| - \sum_{\beta \in \mathcal{V}_p^+} \|\widetilde{h}_k - \widetilde{h}_\beta\| + \epsilon, 0) \\
\mathcal{L}_a &= \max(\sum_{\alpha \in \mathcal{V}_a^+} \|\widetilde{h}_k - \widetilde{h}_\alpha\| - \sum_{\beta \in \mathcal{V}_a^-} \|\widetilde{h}_k - \widetilde{h}_\beta\| + \epsilon, 0)
\end{aligned} \tag{3}$$

where $\| \cdot - \cdot \|$ represents the L2 distance. $\mathcal{V}_p^+$ and $\mathcal{V}_p^-$ represent the graph nodes from $\mathbb{G}$ that connect to v_i with high and low edge-level NSI, respectively. Similarly, the $\mathcal{V}_a^+$ and $\mathcal{V}_a^-$ denote the graph nodes with high and low CAB, respectively. We jointly consider $\mathcal{L}_p$ and $\mathcal{L}_a$ by merging them to one loss function defined as $\mathcal{L}_{\text{GNCC}} = \lambda \mathcal{L}_p + \mu \mathcal{L}_a$ where λ and μ are pre-defined hyperparameters. We finally aggregate all graph node embeddings as $\widetilde{z}_n \in \mathbb{R}^d$ for the m^{th} input photo and transform $\widetilde{z}_m$ to $\hat{y}_m \in \mathbb{R}^K$ that represents the final prediction with K outputs. The output with the maximum value corresponds to the final predicted category. We train $\hat{y}_m$ with the photo label y_m using the cross-entropy loss.

Evaluation

Dataset and Experiment Setup

Dataset. We use publicly available MiniImageNet (Vinyals et al. 2016) and Cifar100 (Krizhevsky and Hinton 2009) as two real-world datasets in our experiments. MiniImageNet is a sub-dataset of ImageNet (Russakovsky et al. 2015) that includes 100 class categories, and 600 images per class category. We randomly split the MiniImageNet dataset with 420 training images and 180 testing images per class. Similarly, the Cifar100 dataset is a sub-dataset of 80-million-tiny-image dataset (Prabhu and Birhane 2020) that includes 100 class categories and 600 images per class category. We split the dataset with 100 testing images and 500 training images.

Crowdsourcing Setup. To generate NSI sets, we randomly sample 20% of the labeled training photos from each dataset. We train 100 deep misclassified photo identifiers from the CDRE module on the remaining 80% of each training set. After training, we use the photos from the NSI set to validate each identifier (defined in CDRE module) and collect all misclassified photos. We then sample photos from all misclassified photos to construct the visual-guided duo relational estimation interface in CDRE by tasking crowd workers to estimate both the NSI and CAB of the photos.

Our tasks are deployed on Amazon MTurk, which randomly selects online crowd workers to answer the tasks regardless of their demographic attributes (e.g., race, gender, age) (Zhang et al. 2021a). We only allow crowd workers who have a 95% or higher Human Intelligence Task (HIT) approval rate to answer our tasks to ensure the quality of the answers. For each task, we randomly generate a verification question (e.g., “what is the first letter of Apple?”) for the crowd worker to avoid crowdsourcing attacks from robot algorithms (Sanchez, Rosas, and Hidalgo 2018). We set the payment to crowd workers well above the requirement from MTurk (Amazon 2022). We follow the IRB protocol approved for this project. In our experiment, we collect 11,487 answers and 10,599 answers from the photos in MiniImageNet and Cifar100, respectively.

Baselines

We compare the performance of SocialCrowd with both state-of-the-art 1) deep learning and crowdsourcing based photo classification baselines: **VGG16** (Simonyan and Zisserman 2014), **DenseNet** (Huang et al. 2017), and **HumanCls** (Peterson et al. 2019); and 2) NSI-aware photo classification models: **DeepBounded** (Sengupta et al. 2018) and **DeepEKL** (Olmo, Sengupta, and Kambhampati 2020).

Evaluation Results

Q1: NSI Performance of SocialCrowd. We first evaluate the performance of SocialCrowd and compared schemes in terms of NSI and classification accuracy. We collect the misclassified photos by all the schemes from the testing set. To obtain the ground-truth NSI label of each class category pair, we invite five well-trained independent social science professionals to manually annotate how negative they think one class category from each category pair is misclassified as the other category. In particular, we create the NSI values 1-5 to indicate the degree of NSI as: 1) totally acceptable; 2) acceptable; 3) neutral; 4) unacceptable; and 5) totally unacceptable. We also apply majority voting on the annotations for each category pair to obtain the final NSI value.

We define the metrics NSI-1 and NSI-5 to evaluate the NSI performance of compared schemes. For a misclassified photo, we retrieve the ground-truth NSI value from the annotated category pair with y and $\hat{y}$ as the first and second categories. We average the retrieved NSI values of all the misclassified photos from the testing set as the NSI-1 score. In addition, we further define NSI-5 to be the NSI value when none of the top 5 most likely labels predicted by a scheme is the same as the photo label. Intuitively, the lower scores of NSI-1 and NSI-5 indicate better NSI performance. We summarize the results in Table 1. We observe that SocialCrowd outperforms all compared schemes with a significant decrease in the NSI value. Such performance gains are mainly attributed to the design of SocialCrowd that accurately estimates the NSI of observed misclassified photos by leveraging crowdsourced human intelligence and explicitly considers the NSI between class categories to reduce the NSI of misclassified photos.

Data	MiniImageNet		Cifar100	
Metric	NSI-1	NSI-5	NSI-1	NSI-5
VGG16	3.06	3.05	3.44	3.41
DenseNet	3.27	3.30	3.01	3.02
HumanCls	3.19	3.56	3.02	2.97
DeepBounded	3.01	3.05	2.98	2.96
DeepEKL	2.74	2.81	2.44	2.50
SocialCrowd	**2.06**	**2.07**	**2.30**	**2.32**

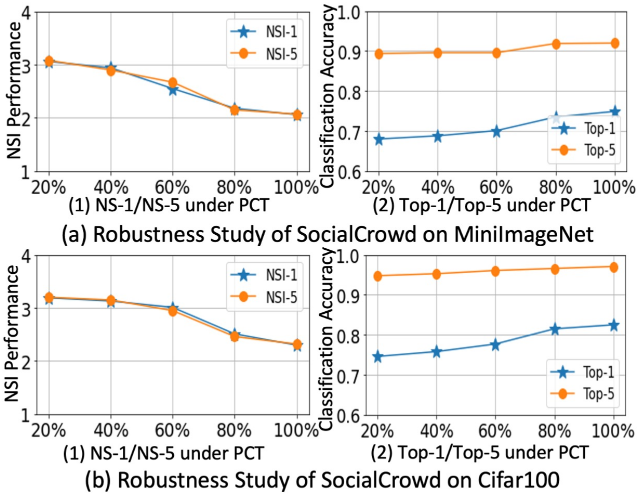
Table 1: NSI Performance

Q2: Photo Classification Performance of SocialCrowd. In addition, we also evaluate the photo classification accuracy of all methods, which is as important as reducing the NSI for classification models. In particular, we evaluate the classification accuracy using the Top-1 and Top-5 metrics that are widely adopted in the computer vision community (Szegedy et al. 2017). The results are shown in Table 2. We observe that SocialCrowd outperforms all compared schemes in terms of the Top-1 metric and outperforms most schemes in terms of the Top-5 metric on both datasets. The reason is that SocialCrowd jointly considers both CAB and NSI of misclassified photos in CVRP to generate discriminative visual embeddings between different graph nodes. The generated graph embeddings are further integrated with the input photo embeddings to improve the classification accuracy of the input photos. However, we also observe the SocialCrowd does not perform as well as VGG16 on classifying photos in terms of the Top-5 metric. One possible reason is that the semantic embeddings and visual embeddings of the graph nodes from CVRP cannot be effectively integrated due to the domain discrepancy between the semantic information (e.g., class category names) and the visual information (e.g., photos from the training and NSI sets).

Q3: Robustness of SocialCrowd: We further study the robustness of SocialCrowd with respect to the key variable: *percentage of crowdsourcing task (PCT)*. The PCT indicates the percentage of crowdsourcing tasks to collect answers

Data	MiniImageNet		Cifar100	
Metric	Top-1	Top-5	Top-1	Top-5
VGG16	0.709	0.890	**0.841**	0.970
DenseNet	0.669	0.897	0.812	0.969
HumanCls	0.687	0.887	0.822	0.969
DeepBounded	0.653	0.866	0.728	0.943
DeepEKL	0.683	0.883	0.785	0.956
SocialCrowd	**0.749**	**0.920**	0.825	**0.971**

Table 2: Classification Accuracy Performance



(a) Robustness Study of SocialCrowd on MiniImageNet

(b) Robustness Study of SocialCrowd on Cifar100

Figure 4: Robustness Study of SocialCrowd

from crowd workers and construct GDRP. We tune PCT from 20% to 100%. The results are shown in Figure 4. We observe that the NSI performance and the classification accuracy improve as PCT increases. The results demonstrate that our crowdsourcing tasks improve the performance of SocialCrowd in terms of both NSI and classification accuracy. However, we also observe the performance of SocialCrowd gradually plateaus especially when the PCT is greater than 60%. One possible reason is that the information obtained from the crowd in the new tasks (e.g., new NSI and CAB scores of misclassified photos) are similar to the existing information embedded in GDRP as the number of crowdsourcing tasks increases.

Q4: Ablation Study of SocialCrowd. Finally, we perform a comprehensive *ablation study* to understand the contributions of important components of SocialCrowd. We create three variants of SocialCrowd by changing its key components: 1) SocialNSI: we do not consider CAB of misclassified photos; 2) SocialCAB: we do not consider NSI of misclassified photos; 3) SocialGraph: we remove the input photo embedding from the concatenation of the graph node embeddings in the GNCC module and aggregate the photo embedding for final photo classification. We summarize the results in Table 3. We observe SocialCrowd outperforms all other variants in terms of all evaluation metrics. The results demonstrate the importance and necessity of the key components of SocialCrowd.

MiniImageNet				
Metric	NSI-1	NSI-5	Top-1	Top-5
SocialNSI	2.28	2.37	0.659	0.885
SocialCAB	2.94	2.98	0.667	0.893
SocialGraph	2.49	2.45	0.680	0.904
SocialCrowd	**2.06**	**2.07**	**0.749**	**0.920**

Cifar100				
Metric	NSI-1	NSI-5	Top-1	Top-5
SocialNSI	2.35	2.37	0.758	0.953
SocialCAB	3.02	3.02	0.801	0.965
SocialGraph	2.54	2.40	0.806	0.966
SocialCrowd	**2.30**	**2.32**	**0.825**	**0.971**

Table 3: Ablation Study of SocialCrowd

Conclusion

This paper presents *SocialCrowd*, a crowd-AI collaborative NSI-aware photo classification framework to address the problem of negative social impact of misclassified photos in large-scale photo classification applications. In particular, we design a novel duo relational NSI-aware graph network to jointly model the NSI and CAB for different class category pairs by exploring the crowdsourced human intelligence. We also develop a context-driven visual relation predictor to efficiently predict the NSI and CAB of unobserved class category pairs. We evaluate SocialCrowd on two large-scale image datasets. Evaluation results show that SocialCrowd significantly outperforms state-of-the-art baseline methods by accurately classifying photos and effectively reducing the NSI of misclassified photos.

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

Amazon. 2022. Amazon Mechanical Turk Pricing. https://www.mturk.com/pricing. Accessed: 2022-08-01.

Balayn, A.; Soilis, P.; Lofi, C.; Yang, J.; and Bozzon, A. 2021. What do You Mean? Interpreting Image Classification with Crowdsourced Concept Extraction and Analysis. In *Proceedings of the Web Conference 2021*, 1937–1948.

Blanco, R.; Halpin, H.; Herzig, D. M.; Mika, P.; Pound, J.; Thompson, H. S.; and Tran Duc, T. 2011. Repeatable and reliable search system evaluation using crowdsourcing. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 923–932.

Borras Jr, S. M.; Moreda, T.; Alonso-Fradejas, A.; and Brent, Z. W. 2018. Converging social justice issues and movements: implications for political actions and research. *Third World Quarterly*, 39(7): 1227–1246.

Chen, H.; and Bu, Y. 2019. Anthropocosmic vision, time, and nature: Reconnecting humanity and nature. *Educational Philosophy and Theory*, 51(11): 1130–1140.

Chen, W.; Liu, Y.; Wang, W.; Bakker, E.; Georgiou, T.; Fieguth, P.; Liu, L.; and Lew, M. S. 2022. Deep Learning for Instance Retrieval: A Survey. arXiv:2101.11282.

Friedman, B.; and Kahn Jr, P. H. 2007. Human values, ethics, and design. In *The human-computer interaction handbook*, 1267–1292. CRC press.

Heim, E.; Roß, T.; Seitel, A.; März, K.; Stieltjes, B.; Eisenmann, M.; Lebert, J.; Metzger, J.; Sommer, G.; Sauter, A. W.; et al. 2018. Large-scale medical image annotation with crowd-powered algorithms. *Journal of Medical Imaging*, 5(3): 034002.

Hettiachchi, D.; Schaekermann, M.; McKinney, T.; and Lease, M. 2021. The Challenge of Variable Effort Crowdsourcing and How Visible Gold Can Help. *arXiv preprint arXiv:2105.09457.*

Hu, J.; Liao, X.; Wang, W.; and Qin, Z. 2021. Detecting Compressed Deepfake Videos in Social Networks Using Frame-Temporality Two-Stream Convolutional Network. *IEEE Transactions on Circuits and Systems for Video Technology.*

Hu, J.; Sun, Z.; Sun, Y.; and Shi, J. 2018. Accumulative image categorization: a personal photo classification method for progressive collection. *Multimedia tools and applications*, 77(24): 32179–32211.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Hui, K.; and Berberich, K. 2017. Transitivity, time consumption, and quality of preference judgments in crowdsourcing. In *European Conference on Information Retrieval*, 239–251. Springer.

Inel, O.; Haralabopoulos, G.; Li, D.; Van Gysel, C.; Szlávik, Z.; Simperl, E.; Kanoulas, E.; and Aroyo, L. 2018. Studying topical relevance with evidence-based crowdsourcing. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 1253–1262.

Korayem, M.; Templeman, R.; Chen, D.; Crandall, D.; and Kapadia, A. 2016. Enhancing lifelogging privacy by detecting screens. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4309–4314.

Kou, Z.; Shang, L.; Zhang, Y.; Duan, S.; and Wang, D. 2022a. Can I only share my eyes? A Web Crowdsourcing based Face Partition Approach Towards Privacy-Aware Face Recognition. In *Proceedings of the ACM Web Conference 2022*, 3611–3622.

Kou, Z.; Zhang, Y.; Zhang, D.; and Wang, D. 2022b. CrowdGraph: A Crowdsourcing Multi-modal Knowledge Graph Approach to Explainable Fauxtography Detection. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–28.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*.

Lonn, S.; Radeva, P.; and Dimiccoli, M. 2019. Smartphone picture organization: A hierarchical approach. *Computer Vision and Image Understanding*, 187: 102789.

Olmo, A.; Sengupta, S.; and Kambhampati, S. 2020. Not all failure modes are created equal: Training deep neural networks for explicable (mis) classification. *arXiv preprint arXiv:2006.14841.*

Perera, P.; Oza, P.; and Patel, V. M. 2021. One-class classification: A survey. *arXiv preprint arXiv:2101.03064.*

Peterson, J. C.; Battleday, R. M.; Griffiths, T. L.; and Russakovsky, O. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9617–9626.

Prabhu, V. U.; and Birhane, A. 2020. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923.*

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Sanchez, L.; Rosas, E.; and Hidalgo, N. 2018. Crowdsourcing under attack: Detecting malicious behaviors in Waze. In *IFIP International Conference on Trust Management*, 91–106. Springer.

Saralioglu, E.; and Gungor, O. 2019. Use of crowdsourcing in evaluating post-classification accuracy. *European Journal of Remote Sensing*, 52(sup1): 137–147.

Savenkov, D.; Weitzner, S.; and Agichtein, E. 2016. Crowdsourcing for (almost) real-time question answering. In *Proceedings of the Workshop on Human-Computer Question Answering*, 8–14.

Sengupta, S. 2020. *The What, When, and How of Strategic Movement in Adversarial Settings: A Syncretic View of AI and Security*. Ph.D. thesis, Arizona State University.

Sengupta, S.; Dudley, A.; Chakraborti, T.; and Kambhampati, S. 2018. An investigation of bounded misclassification for operational security of deep neural networks. In *AAAI Workshop of Engineering Dependable and Secure Maching Learning Systems*.

Shi, X.; Xu, L.; Wang, P.; Gao, Y.; Jian, H.; and Liu, W. 2020. Beyond the attention: Distinguish the discriminative and confusable features for fine-grained image classification. In *Proceedings of the 28th ACM International Conference on Multimedia*, 601–609.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29: 3630–3638.

Wang, M.; and Deng, W. 2019. Mitigate bias in face recognition using skewness-aware reinforcement learning. *arXiv preprint arXiv:1911.10692*.

Wehrli, S.; Hertweck, C.; Amirian, M.; Glüge, S.; and Stadelmann, T. 2021. Bias, awareness, and ignorance in deep-learning-based face recognition. *AI and Ethics*, 1–14.

Zhang, D. Y.; Huang, Y.; Zhang, Y.; and Wang, D. 2020. Crowd-assisted disaster scene assessment with human-ai interactive attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2717–2724.

Zhang, S.; Yao, Y.; Xu, F.; Tong, H.; Yan, X.; and Lu, J. 2019. Hashtag recommendation for photo sharing services. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5805–5812.

Zhang, Y.; Shang, L.; Zong, R.; Wang, Z.; Kou, Z.; and Wang, D. 2021a. StreamCollab: A Streaming Crowd-AI Collaborative System to Smart Urban Infrastructure Monitoring in Social Sensing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 179–190.

Zhang, Y.; Zong, R.; Kou, Z.; Shang, L.; and Wang, D. 2021b. Collablearn: An uncertainty-aware crowd-ai collaboration system for cultural heritage damage assessment. *IEEE Transactions on Computational Social Systems*.

Zhang, Y.; Zong, R.; Kou, Z.; Shang, L.; and Wang, D. 2022a. CrowdNAS: A Crowd-guided Neural Architecture Searching Approach to Disaster Damage Assessment. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–29.

Zhang, Y.; Zong, R.; Shang, L.; Kou, Z.; Zeng, H.; and Wang, D. 2022b. 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.

Zhang, Z.; and Sabuncu, M. R. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*.

Zuccon, G.; Leelanupab, T.; Whiting, S.; Jose, J.; and Azzopardi, L. 2011. Crowdsourcing interactions. *CSDM'11*, 35.