Disaster Image Classification Using Pre-trained Transformer and Contrastive Learning Models

1st Soudabeh Taghian Dinani
Department of Computer Science
Kansas State University
Manhattan, Kansas, USA
soudabehtaghian@ksu.edu

2nd Doina Caragea
Department of Computer Science
Kansas State University
Manhattan, Kansas, USA
dcaragea@ksu.edu

***Abstract*—Natural disasters can have devastating consequences for communities, causing loss of life and significant economic damage. To mitigate these impacts, it is crucial to quickly and accurately identify situational awareness and actionable information useful for disaster relief and response organizations. In this paper, we study the use of advanced transformer and contrastive learning models for disaster image classification in a humanitarian context, with focus on state-of-the-art pre-trained vision transformers such as ViT, CSWin and a state-of-the-art pre-trained contrastive learning model, CLIP. We evaluate the performance of these models across various disaster scenarios, including in-domain and cross-domain settings, as well as few-shot learning and zero-shot learning settings. Our results show that the CLIP model outperforms the two transformer models (ViT and CSWin) and also ConvNeXts, a competitive CNN-based model resembling transformers, in all the settings. By improving the performance of disaster image classification, our work can contribute to the goal of reducing the number of deaths and economic losses caused by disasters, as well as helping to decrease the number of people affected by these events.**

***Index Terms*—disaster image classification, deep learning, transformers, ViT, CSWin, contrastive learning, CLIP, ConvNeXts**

I. Introduction

Nowadays, smartphones and the internet are accessible to a majority of people and, as a result, social media platforms, such as Twitter, have become a quick and popular way of communicating and sharing information during various types of crisis situations [1]. This is especially true during disaster events, as 911 and other emergency lines rapidly become overwhelmed by the large volume of calls from people in need of help [21]. Thanks to the fast spread of news via social media, relief and response during disasters can be accelerated using this vital medium [1]. In fact, social media can disseminate disaster-related news and updates much faster and to a broader audience as compared to traditional media. According to a report by Middle East Eye, several trapped victims were rescued during the 2023 earthquake in Turkey after posting "locations and images" on social media [17]. For example, a man tweeted that he and his family were buried under rubble in Hatay. Within an hour of the post, the tweet was shared thousands of times and the family was rescued by nearby residents. Similarly, another man posted his location and a plea for help "Please help! We are under debris. There are many people here." One hour later, the man reported that "he and his mother had been rescued, but his father was not as fortunate." These examples demonstrate how social media can be used to quickly and effectively communicate information during a disaster, allowing people to share urgent needs with a wide audience and receive help [17].

Information posted on social media, such as requests for help and support, damage reports or updates on relief efforts, can be utilised by humanitarian organizations to facilitate post-disaster rescue and relief operations, guide the allocation of resources, and improve the real-time response overall [4]. Although such information is valuable, being posted in real-time by eyewitnesses of disasters, the amount of information posted is extremely large, and can be noisy, repetitive or even irrelevant. In fact, as soon as a disaster emerges, some people start to sympathize with the victims, express thanks to rescue organisations, repost the same content and so forth [37]. Therefore, it is of great importance to effectively filter and prioritise relevant and informative content. However, manually filtering relevant information is a hard task due to the large volume and high velocity of data, and thus, automatic filtering techniques are required [42].

While research on textual analysis of crisis-related tweets has progressed significantly, research on image analysis of such tweets has not received as much attention. However, recent studies have explored the use of convolutional neural networks (CNNs) [2], [6], [18] to classify disaster images into multiple categories, such as disaster types, image informativeness, humanitarian categories, and damage severity. Nevertheless, more recently, transformer models like the Vision Transformer (ViT) [13] have demonstrated promising results, surpassing CNNs in many computer vision scenarios. For instance, a recent study [50] used a Tokens-to-Token ViT to classify cervical cancer smear cell images. Another study [38] employed a self-supervised pre-trained Swin Transformer for classification and segmentation of land cover, and yet another study [5] used a transformer-based model, LPViT, for defect detection and classification of printed circuit boards (PCBs).

Despite the success of transformer and contrastive learning models in various application domains, their potential for disaster image classification, especially in the case of contrastive learning, has not been much explored yet. Our study aims to

investigate the feasibility of utilizing such models for this task. Given the scarcity of images in the early stages of a disaster, we plan to leverage the strengths of transformer/contrastive learning models in experimental settings such as zero-shot [43] few-shot [45] and also cross-domain settings [47], which have been already extensively studied in the literature for other application domains [9], [16], [19], [22], [32], [35], [40], [41], [44], [46], [49]. Specifically, we investigate the performance of transformer variants, including ViT [13], CSWin Transformer [12], and also a contrastive learning model, the image encoder of CLIP (Contrastive Language-Image Pre-training) [36], by comparison with ConvNetXt (a modern CNN architecture designed to resemble transformers), in the context of disaster image classification.

The main contributions of this study are as follows:

- We use pre-trained transformer models (ViT, CSWin), CLIP (the image encoder based on ViT) and ConvNeXts (a CNN-based model which resembles transformers) for disaster image classification. To the best of our knowledge, our study is the first one to use contrastive learning for disaster image classification.
- We study the pre-trained CLIP model on disaster image classification in a zero-shot transfer learning setting.
- We also study the models in a few-shot learning setting (with 1/5/10/20 instances per class, respectively) and compare the results of the few-shot models with those of the CLIP model in zero-shot setting, which can be seen as a lower-bound, as well as with the results obtained through fine-tuning the pre-trained models using all the data available, which can be seen as an upper-bound.
- We explored CLIP in several other experimental settings, including *in-domain transfer* and *cross-domain transfer*, which further involved two sub-settings of *one-versus-one* and *all-but-one*.
- Our results show that CLIP exhibits strong generalization capability across different disasters and consistently achieves the best overall performance in various settings. These findings suggest that CLIP is a strong candidate for disaster image classification tasks.

Most importantly, the results of this study suggest that our research can contribute to the development of effective solutions for disaster management and response.

II. Related work

In this section, we review related work on CNN models used for disaster image classification as well as work on recent vision transformers and contrastive learning models.

A. Disaster Image Classification

While research on textual analysis of crisis-related tweets has witnessed remarkable progress, research on image analysis of such tweets has remained comparatively under-explored [2]. Inspired by successes of Deep Learning (DL) in the field of image classification, some studies have been conducted on social media disaster images [2], [4], [6], [11], [18], [25].

For instance, Alam et al. [2] fine-tuned convolutional neural networks (CNNs), including ResNet18, ResNet50, ResNet101, VGG16, DenseNet, SqueezeNet, MobileNet, and EfficientNet models, using disaster images for multiple tasks (disaster types, image informativeness, humanitarian categories and damage severity), and showed promising results. In another study, Banerjee et al. [6] used GAN models to generate disaster images to augment existing datasets, and fine-tuned CNN models, such as VGG19, ResNet18, and DenseNet121, to classify different types of real-world disaster images, with ResNet18 proving to be the most effective. Furthermore, Irwansyah et al. [18] utilized models such as PSPNet and UNet with ResNet18 and ResNet50 backbones to classify satellite disaster images into four damage classes. In [29], authors utilized average pooling to compress initial feature maps (from the convolution layer) into three distinct strips: 'row strip', 'column strip', and 'channel strip'. Applying separate attention weighting to each strip, they integrated them, resulting in the Triple-Strip Attention Mechanism (TSAM), targeted at capturing often overlooked global features by convolutions. The method was employed for both disaster image classification, distinguishing between landslide and non-landslide images, and segmentation tasks, has been shown to effectively handle landslide and flood disaster image segmentation. Collectively, these studies demonstrate the effectiveness of CNN models in disaster image classification, showcasing their wide applicability across different image-based tasks and contexts.

However, most recently, transformer models, such as Vision Transformer (ViT) [13], have emerged as a powerful tool for computer vision tasks, including classification and segmentation, and have surpassed CNNs in performance [13], [38]. These models have been applied to various application domains with very promising results [15], [20], [23], [28], as discussed in the next subsection. Another well-known contemporary model, CLIP (Contrastive Language-Image Pre-training) [36], is a contrastive learning multimodal model. Trained on a large image-text collection, CLIP demonstrates remarkable transferability, effectively accommodating diverse tasks via fine-tuning. Its distinctive zero-shot and few-shot capabilities position CLIP as a well-suited choice for disaster response settings characterized by limited initial labeled data. However, to date, CLIP has not been applied to disaster image classification, and transformer-based models remain unexplored in zero-shot, few-shot scenarios, and humanitarian task.

B. Vision Transformers and Contrastive Learning

The introduction of vision transformers, such as ViT [13], has brought about a paradigm shift in image analysis, particularly in the realm of image classification. Transformer-based models have established a strong presence in various image analysis domains, largely due to their superior performance compared to state-of-the-art CNNs [13], making it increasingly difficult for CNNs to remain competitive.

In an effort to keep up with the superior performance of transformers, a "modernized" version of ResNet, known as

ConvNeXts, was developed in [27]. ConvNeXts was designed to resemble transformers and produced very competitive results in terms of accuracy, scalability, and robustness. At the same time, ViT models pre-trained on large datasets have resulted in outstanding performance when transferred to smaller downstream datasets [13], and few-shot results of ViT pre-trained on ImageNet have also been promising. For our application domain, transformer-based models pre-trained on large datasets are ideal since obtaining reliable labels requires significant time and effort, and the availability of labeled data during a disaster is limited, especially at the beginning.

Given the initial success of the ViT model, other improved versions of transformers have also emerged. For example, Swin Transformer [26], a hierarchical transformer whose representation is computed with shifted windows, was introduced and was shown to surpass the counterpart ViT on the ImageNet-1K dataset. A further developed version of Swin Transformer was presented in [12], called CSWin Transformer, in which self-attention was performed in horizontal and vertical stripes, in parallel, forming a Cross-Shaped Window self-attention. CLIP (Contrastive Language-Image Pre-training) is a popular contrastive learning multimodal model [36], which consists of an image encoder (e.g., ViT transformer architecture) and a text encoder (e.g., BERT transformer [10]) pre-trained together on image-caption pairs using a constrastive loss that aims to produce similar representations for an image and its corresponding caption. Since CLIP was pre-trained on a large scale dataset consisting of 400 million image–text pairs [36], it has outstanding transfer capabilities and has been successfully fine-tuned to other tasks. Moreover, CLIP has also shown good zero-shot and few-shot capabilities, which makes it an ideal candidate for a disaster response setting, as labeled data is generally not available in the beginning of a disaster.

Considering the superior performance of transformer-based models in image classification tasks and their ability to transfer well to downstream tasks with datasets of small or medium size, they are a suitable option for disaster image classification tasks with limited labeled data. To this end, we study the pre-trained CLIP [36], specifically its image-encoder component, by comparison with other well-known pre-trained transformers, such as ViT [13] and CSWin [12], for disaster image classification in a variety of settings. In addition, we use ConvNeXt [27], a CNN model that resembles hierarchical vision Transformers, as a baseline CNN.

III. Approaches

In this section, we briefly review the approaches used in this study, including ViT, ConvNeXts, CSWin, and CLIP.

A. Vision Transformer (ViT)

ViT is a vision model based on the transformer architecture. The transformer was originally designed for natural language processing (NLP) and resulted in performance improvements for many NLP tasks [13]. The transformer utilizes a self-attention mechanism to capture long-range dependencies and to produce a global feature representation of the input [13]. In ViT, the image is split into patches of fixed-size, which are processed by a transformer encoder after being linearly embedded [13]. ViT adds a unique CLS token to the input to obtain a global representation for classification tasks [7], [13].

B. Cross-Shaped Window Transformer (CSWin)

CSWin is another vision model based on the transformer architecture. As opposed to ViT, which was targeted at image classification, CSWin was developed for general-purpose vision tasks [12]. CSWin's main component is a Self-Attention module, which executes self-attention calculations in parallel by dividing multi-heads into parallel groups and processing the horizontal and vertical stripes simultaneously. This technique results in an efficient enlargement of the attention area for each token in a transformer block. Furthermore, by adjusting the stripe width according to the network depth, the attention area can be further increased with minimal additional computational cost. To enhance its capabilities, CSWin Transformer incorporates the Locally-enhanced Positional Encoding (LePE), which can handle different input resolutions [12].

C. ConvNeXts

ConvNeXt is a pure convolution-based model which was designed by "modernizing" the ResNet architecture to resemble a hierarchical vision transformer [27]. Its architecture features depthwise convolutions and an inverted bottleneck design, which reduces Floating Point Operations (FLOPs) and enhances performance. Additionally, the incorporation of larger kernel sizes improves the network's global receptive field and accuracy. ConvNeXt has been shown to rival transformers in terms of accuracy and scalability, while maintaining the simplicity and efficiency of standard CNNs [27].

D. Contrastive Language-Image Pre-Training

CLIP is a multimodal architecture that enables joint training of NLP and computer vision encoders, by utilizing a contrastive loss. The contrastive loss encourages the model to produce similar representations for semantically related image-text pairs, while dissimilar pairs are pushed further apart in the joint image-text representation space [36]. CLIP has been trained in a self-supervised manner on a dataset of 400 million image-caption pairs, with ResNet-50/ViT, respectively, as image encoders and BERT [10] as a transformer-based text encoder. We used ViT as the image encoder in this work. Therefore, when referring to the use of the CLIP model in this paper, we specifically mean CLIP with ViT as the image encoder (and generally refer to this model simply as CLIP). Unlike traditional approaches that rely on pre-defined object categories, CLIP learns to identify objects based on textual descriptions. This enables CLIP to excel in tasks such as zero-shot and few-shot image classification. To perform zero-shot image classification, the input image is paired one-by-one with the textual names of the categories included in the classification task (e.g., text such as "A photo of a CLASS-NAME"), and the pair is embedded with the pre-trained model. The encoded image is compared to all encoded categories to determine the category with the highest similarity [20], [36].

TABLE I: Data distribution of the CrisisMMD dataset for the Informativeness task

Dataset	Event	Non-inf.	Inf.	Total
D0	California Fire	604	984	1588
D1	Hurricane Harvey	1982	2461	4443
D2	Hurricane Irma	2303	2222	4525
D3	Hurricane Maria	2330	2232	4562
D4	Iraq Iran Earthquake	200	400	600
D5	Mexico Earthquake	541	841	1382
D6	Sri Lanka Floods	773	252	1025

TABLE II: Data distribution of the CrisisMMD dataset for the modified Humanitarian task.

Dataset	Affected	Damage	Effort	Other	Total
D0	38	601	205	139	983
D1	192	1032	666	570	2460
D2	62	887	366	907	2222
D3	153	924	439	711	2227
D4	101	182	44	72	399
D5	75	210	446	104	835
D6	51	101	69	31	252

IV. Experimental setup

In this section, we provide an overview of the dataset used in our experiments, the experimental setup, and the evaluation metrics used to evaluate the models.

A. Dataset

We have utilized the CrisisMMD dataset [3] for our study. This dataset consists of tweets containing images and texts from seven natural disasters, namely Hurricane Irma, Hurricane Harvey, Hurricane Maria, California Wildfires, Mexico Earthquake, Iraq-Iran Earthquake, and Sri Lanka Floods. We have focused on two different tasks available in the CrisisMMD dataset, namely Informativeness task (i.e., predicting if an image is informative with respect to a particular disaster or not informative) and Humanitarian task (i.e., predicting the humanitarian category associated with an informative image).

The Informativeness task is a binary classification task, specifically Informative (*Inf.*) and Non-informative (*Non-inf.*) with respect to a disaster of interest. The original Humanitarian task in CrisisMMD consists of eight categories, including affected-individuals, injured-or-dead-people, missing-or-found-people, infrastructure-and-utility-damage, vehicle-damage, rescue-volunteering-or-donation-effort, not-relevant-or-cant-judge, and other-relevant-information. We omitted the missing-or-found-people category due to its relatively small number of instances. We refer to the rescue-volunteering-or-donation-effort as *Effort*. Moreover, we combined the affected-individuals and injured-or-dead-people categories into one category called affected-injured-or-dead-people or simply *Affected*. Similarly, we combined infrastructure-and-utility-damage and vehicle-damage categories into one category called *Damage*. Finally, we merged the not-relevant-or-cant-judge and other-relevant-information categories into a single category called *Other*. Thus, the final dataset consists of images in four categories: *Affected*, *Damage*, *Effort* and *Other*. The distribution of datasets for the Informativeness and Humanitarian tasks are presented in Table I and II, respectively.

Data Splits: We randomly split each dataset into three subsets: training (70%), development (10%), and test (20%). We used three different random seeds for splitting, resulting in three different splits for the training, development, and test subsets. Each experiment was run on the three splits, and the results are reported in terms of averages over the three runs.

Note: In terms of utilizing the CrisisMMD dataset, there exist considerable variations in how researchers split the datasets, determine the number of classes for each task, and whether they present results based on the entire merged dataset or individual subsets. While Ofli *et al.* [34] established standard benchmark splits for CrisisMMD tasks when utilizing the dataset for multimodal approaches, there is an observed discrepancy in the adoption of CrisisMMD dataset across studies, particularly when focusing on a single modality (text or image) for classification purposes. Notably, when only image modality is used for classification, Alam *et al.* [4] provided benchmark splits for the combined set of all seven datasets, as opposed to individual subsets. Moreover, although humanitarian task has 8 categories, some investigations employ a modified 4-class version of CrisisMMD for experimentation, as exemplified by [4], while some others use a 5-class version, as evidenced in works such as [24], [34], [39]. Also some works report results for all the crisis events combined [4], [14], [30], while some other studies report results for different crisis events separately. In the latter case, specific studies use all seven events [31], [48], while others narrow their focus to a selected subset of these events [8], [33]. Thus, a comparison of prior approaches that have used the CrisisMMD dataset is hardly possible. Given this and our focus on the performance of pretrained transformers and contrastive learning CLIP model by comparison with CNNs for image classification in a variety of low-data settings, we have chosen to use classes that are well represented in the dataset (sometimes obtained by merging two of the original classes as described above). As we work with independent events, we also chose to split the data per event according to standard practices to create three random splits. This allows us to report average results over three splits (and thus capture variation in the performance of the models).

B. Evaluation Metrics

In all experiments, we chose weighted precision, weighted recall, and weighted F1-score as the evaluation metrics. We present only the F1-score values in the main result tables due to limited space.

C. Experimental Settings

We aim to study the effectiveness of the models considered in realistic scenarios, where little to no labeled data is generally available for an emergent disaster. Specifically, we focus on zero-shot, few-shot and in-domain/cross-domain transfer learning settings. Our baseline setting corresponds to the traditional supervised setting, where labeled data is

available to train a model for a specific disaster of interest. Among the four models studied, ConvNetXt (a CNN model) is used as a baseline for the transformer-based models (ViT and CSWin) and the contrastive learning model, CLIP. Note that we generally use the image encoder of CLIP, specifically ViT, in our experiments, although we refer to the model as CLIP. More details for the different settings included in our experimental design are provided below.

Event-specific supervised (baseline): In this setting, we train event-specific models (ConvNetXt, ViT, CSWin, and CLIP) for each of the seven datasets (events) in CrisisMMD separately. Given an event dataset D, we use the training subset of the dataset D to train (fine-tune) the pre-trained models for the Informativeness and Humanitarian tasks, respectively. We use the validation subset of D to select hyper-parameters, and the test subset of D to evaluate the performance of the model. The term "event-specific" denotes the use of a single dataset for both the training and test subsets.

Event-specific zero-shot: In the zero-shot setting, we evaluate the pre-trained CLIP model by pairing each input test image with text representing category names, one at a time (e.g., "A photo of affected, injured or dead people"). We select the category corresponding to the text with the highest similarity with the input image. The test subset of an event dataset D is used to evaluate the performance for that event. Note that we only use CLIP in the event-specific zero-shot setting as the other models do not have the ability to assign specific labels to instances without any additional fine-tuning.

Event-specific few-shot: In the few-shot setting, we use 1/5/10/20 instances per class to fine-tune models (ConvNetXt, ViT, CSWin, and CLIP) for a particular event D. The instances used are randomly selected from the training set of that event D in a cumulative fashion (by adding more instances to the previous subset). The development and test subsets of D are used to select hyper-parameters and to evaluate the performance of the models, respectively.

Based on the results of the experiments in the event-specific supervised and few-shot settings (which will be discussed in Section V), we selected CLIP as a best performing model on our disaster-related tweet classification tasks and used it for additional experimentation with in-domain and cross-domain transfer settings, as described below.

In-domain transfer: In this setting, we use CLIP models trained on prior events of a specific disaster type (called *source* events) to predict data from an emergent event of the same type (called *target* event). The fine-tuning of the models is done on the training data from the source events. The hyper-parameter selection and model evaluation are performed on the development and test subsets of the target event, respectively. There are two types of disasters with multiple events in the CrisisMMD dataset, specifically, hurricanes and earthquakes. Thus, we perform in-domain experiments using these two types of disasters independently. In each in-domain experiment, one event of a specific type is used as target, while the other event(s) are used as source.

Cross-domain transfer: In the cross-domain setting, we train models on source events of some type(s) and evaluate the models on events of potentially different type. More specifically, we consider two sub-settings here. First, we consider a **one-versus-one** setting in which the model is trained on a source event of a particular disaster type (e.g., hurricane) and evaluated on a target event of a different type (e.g., flood). This setting may be useful in situations where the source and target events happen around the same time or at the same location, or if a prior event of the same type as the target event is not available. Second, we consider the **all-but-one** setting (a.k.a., **leave-one-out** setting) where all available events (regardless of their type) are used as source to train a model for a left-out target event. This setting is helpful in understanding if more training data from a variety of events is better than a smaller amount of training data from more similar events to a specific target event. The cross-domain models are fine-tuned on the training data from the source events. The hyper-parameter selection and model evaluation are performed on the development and test subsets of the target event, respectively.

D. Research Questions

Our experiments aim to answer the following research questions (RQ):

(RQ1) Among the four models considered, which model yields the best results for disaster image classification in the event-specific supervised and few-shot settings?

(RQ2) How do the models perform in the few-shot setting by comparison with the supervised learning setting?

(RQ3) How does the CLIP model perform in-domain and cross-domain by comparison with few-shot/supervised settings?

(RQ4) What setting leads to the best results for an emergent disaster for which little or no labeled data is available?

E. Hyper-parameters

A simple data augmentation approach was applied to the training set of each model. Images were resized to 225 × 225 and then randomly cropped to 224 × 224. Additionally, images were randomly flipped in the horizontal direction and normalized. For the test and development sets, images were only resized to 224 × 224 and normalized. The specific pre-trained models used in the experiments are as follows: "facebook/convnext-tiny-224" for ConvNeXts, "google/vit-base-patch16-224-in21k" for ViT, "CSWin-Tiny (CSWin_64_12211_tiny_224)" for CSWin and "ViT-B/32" for CLIP. The last layer in each pre-trained model was removed and replaced with a linear classifier (along with dropout) that had the output size equal to the number of categories in a particular task (specifically, 2 for the Informativeness task and 4 for the Humanitarian task). All layers were frozen except for the last layer (corresponding to the linear classifier), whose parameters were randomly initialized and trained from scratch. We used the Adam optimizer, a batch size of 32 and a maximum of 50 epochs for all experiments. The exact number of epochs for each experiment was determined separately based on the development set. The best model according to the

development set was used to estimate the final performance on the corresponding test set.

V. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we fist present the results of the experiments that we performed, followed by a thorough discussion of the results and error analysis.

A. Results

Table III shows the results of the comparison between the ConvNeXts, ViT, CSWin and CLIP models in event-specific supervised and few-shot settings for both Informativeness task and Humanitarian task. This table also shows the zero-shot results for the CLIP model. The table shows F1 scores for each event and also averages over all events for the supervised setting. In the few-shot setting, the F1 scores corresponding to k instances (shots) per class ($k = 0$, 1, 5, 10 and 20, respectively) are averaged over all events in the dataset due to space limitations. The supervised setting is used as a baseline for the few-shot setting, while the ConvNeXts is used as a baseline for the transformer/CLIP models. The results of the baseline ConvNeXts are shown in white, while the increment and decrement in the F1-score of other models included in the comparison are indicated with shades of green and red, respectively; the darker the shade, the higher the increment /decrement. The best F1-score in each row is **bold-faced**.

Table IV presents additional results for CLIP, the best performing model overall. Specifically, few shot results (with 20 instances per class), and in-domain and cross-domain transfer results are shown for each event separately, for both Informativeness and Humanitarian tasks, by comparison with the results of the supervised setting, which is the baseline setting. In this table, the baseline results in the supervised setting are shown in white, while the increment and decrement in the F1-score of other settings compared to the baseline are indicated with the shades of green and red, respectively. As for Table III, the darker the shade, the higher the increment/decrement, and the best F1-score in a row is **bold-faced**.

B. Discussion

In what follows, we will use the results in Tables III and IV to answer our research questions in Section IV-D.

(RQ1) *Among the four models considered, which model yields the best results for disaster image classification in the event-specific supervised and few-shot settings?* The results in Table III indicate that for both tasks, CLIP generally outperforms the other models in terms of F1-score by a large margin, in both supervised and few-shot settings. For example, in the supervised setting, for the informativeness task, CLIP's average F1-score over the seven events in the dataset is 86.571, while the average F1-score of the ConvNetXts baseline is 82.292. Similarly, for the humanitarian task, CLIP's average F1-score is 81.805, while ConvNetXts's average F1-score is 78.198. Similar results can be observed in the few-shot setting, where CLIP consistently and significantly outperforms ConvNetXts for both tasks. Notably, CLIP improves ConvNetXts's average F1-score by almost 10% for the informativeness task, when 5 shots are used. It is also remarkable to see that CLIP has an average F1-score of approximately 50% in the zero-shot setting for both tasks, although as expected fine-tuning the model using a small number of instances improves the performance. We believe that the superior performance of CLIP, followed by ViT as the second-best model, can be attributed to several factors, including the use of the ViT

TABLE III: F1-score of event-specific models in supervised, zero-shot and few-shot settings for the Informativeness and Humanitarian tasks, respectively. The F1-scores for the event-specific supervised setting are reported per event, while the F1-scores for the event-specific zero-shot and few-shot settings are averaged over all events in the dataset. Models compared include ConvNeXts (baseline CNN model) and ViT, CSWin and CLIP (transformer-based models). The shades of green and red indicate the increment and decrement in the F1-scores of transformer models as compared to ConvNeXts. The darker the shade, the higher the percentage of increment/decrement. The F1-scores of the ConvNeXts model are shown in white, while the best F1-score in each row is **bold-faced**.

Event-specific supervised setting				
Informativeness Task (F1-scores per event)				
Dataset	ConvNeXts	CSWin	ViT	CLIP
D0 (Fire)	85.335	82.962	86.269	**86.832**
D1 (Hurricane)	82.741	81.062	85.105	**87.431**
D2 (Hurricane)	79.808	78.763	80.622	**84.853**
D3 (Hurricane)	79.606	79.824	81.216	**84.745**
D4 (Earthquake)	81.236	80.915	81.896	**84.596**
D5 (Earthquake)	81.500	81.949	80.051	**86.495**
D6 (Flood)	85.819	86.392	88.412	**91.047**
Average	82.292	81.695	83.367	**86.571**
Humanitarian Task (F1-scores per event)				
Dataset	ConvNeXts	CSWin	ViT	CLIP
D0 (Fire)	78.999	77.698	72.779	**81.047**
D1 (Hurricane)	80.915	79.583	80.371	**82.772**
D2 (Hurricane)	83.832	84.960	84.912	**85.097**
D3 (Hurricane)	84.214	84.027	84.219	**86.072**
D4 (Earthquake)	76.069	77.278	77.786	**80.531**
D5 (Earthquake)	72.473	74.924	74.978	**80.870**
D6 (Flood)	70.887	64.374	75.052	**76.243**
Average	78.198	77.549	78.585	**81.805**
Event-specific zero-shot and few-shot settings				
Informativeness Task (F1-scores averaged over all events)				
#Shots	ConvNeXts	CSWin	ViT	CLIP
0	-	-	-	50.816
1	55.791	50.111	61.458	**61.694**
5	66.670	66.070	69.805	**76.418**
10	70.292	70.252	72.416	**78.972**
20	73.077	73.772	76.476	**81.713**
All (supervised)	82.292	81.695	83.367	**86.571**
Humanitarian Task (F1-scores averaged over all events)				
#Shots	ConvNeXts	CSWin	ViT	CLIP
0	-	-	-	49.688
1	47.508	**58.285**	49.262	47.855
5	63.792	66.746	61.694	**67.044**
10	70.208	69.205	70.179	**74.594**
20	73.958	72.241	73.951	**76.805**
All (supervised)	78.198	77.549	78.585	**81.805**

TABLE IV: CLIP F1-scores for Informativeness and Humanitarian tasks in supervised, few-shot, in-domain/cross-domain transfer settings. The shades of green and red indicate the increment and decrement in the F1-scores of the few-shot, in-domain, cross-domain transfer settings as compared to the supervised setting (baseline, white). The darker the shade, the higher the percentage of increment/decrement. The best overall not-supervised F1-score for each event is **bold-faced**.

Event-specific supervised settings			
Source	Target	Inf.	Hum.
D0 (Fire)	D0 (Fire)	86.832	81.047
D1 (Hurricane)	D1 (Hurricane)	87.431	82.772
D2 (Hurricane)	D2 (Hurricane)	84.853	85.097
D3 (Hurricane)	D3 (Hurricane)	84.745	86.072
D4 (Earthquake)	D4 (Earthquake)	84.596	80.531
D5 (Earthquake)	D5 (Earthquake)	86.495	80.870
D6 (Flood)	D6 (Flood)	91.047	76.243
Event-specific few-shot setting			
#Shots/Source	Target	Inf.	Hum.
20 of D0 (Fire)	D0 (Fire)	82.811	**78.724**
20 of D1 (Hurricane)	D1 (Hurricane)	81.854	74.186
20 of D2 (Hurricane)	D2 (Hurricane)	79.650	80.737
20 of D3 (Hurricane)	D3 (Hurricane)	78.782	83.598
20 of D4 (Earthquake)	D4 (Earthquake)	80.531	81.297
20 of D5 (Earthquake)	D5 (Earthquake)	81.609	**73.097**
20 of D6 (Flood)	D6 (Flood)	86.751	65.997
In-domain transfer settings (cross-event)			
Source	Target	Inf.	Hum.
D2 + D3 (Hurricane)	D1 (Hurricane)	85.922	80.980
D1 + D3 (Hurricane)	D2 (Hurricane)	**85.151**	85.477
D1 + D2 (Hurricane)	D3 (Hurricane)	**84.540**	**84.576**
D5 (Earthquake)	D4 (Earthquake)	**85.992**	63.821
D4 (Earthquake)	D5 (Earthquake)	82.039	34.563
Cross-domain transfer settings (one-versus-one)			
Source	Target	Inf.	Hum.
D3 (Hurricane)	D0 (Fire)	84.423	74.287
D5 (Earthquake)	D0 (Fire)	83.430	76.283
D6 (Flood)	D0 (Fire)	61.113	58.872
D0 (Fire)	D3 (Hurricane)	78.626	78.804
D5 (Earthquake)	D3 (Hurricane)	79.462	80.295
D6 (Flood)	D3 (Hurricane)	75.675	60.798
D0 (Fire)	D5 (Earthquake)	78.852	56.731
D3 (Hurricane)	D5 (Earthquake)	83.734	61.605
D6 (Flood)	D5 (Earthquake)	67.489	38.179
D0 (Fire)	D6 (Flood)	88.401	61.457
D3 (Hurricane)	D6 (Flood)	**90.696**	75.031
D5 (Earthquake)	D6 (Flood)	89.640	60.317
Cross-domain transfer settings (all-but-one)			
Source	Target	Inf.	Hum.
All but D0	D0 (Fire)	**85.841**	76.415
All but D1	D1 (Hurricane)	**86.822**	**81.524**
All but D2	D2 (Hurricane)	85.000	**86.294**
All but D3	D3 (Hurricane)	84.138	83.765
All but D4	D4 (Earthquake)	83.366	**82.810**
All but D5	D5 (Earthquake)	**86.936**	65.441
All but D6	D6 (Flood)	90.264	**79.034**

model itself as the image encoder of CLIP, and also the fact that CLIP was pre-trained on a large and diverse corpus of multimodal image-caption pairs, thus enabling the model to identify more subtle image concepts that the text describes and also to generalize well to different domains and tasks.

When analyzing the results of the two transformer models (CSWin and ViT), we can see that ViT is overall better than ConvNetXts, but worse than CLIP. However, ViT's improvement over ConvNetXts is more significant for the informativeness task, while the results of the two models are somewhat comparable for the humanitarian task. Finally, CSWin seems to have the worse overall performance on the two disaster image classification tasks in our study.

(RQ2) *How do the models perform in the few-shot setting by comparison with the supervised learning setting?* Average F1 scores over the seven events in the dataset can be seen in Table III (lower part) for all models considered, in the supervised setting (All) and also in the few-shot setting (where 1, 5, 10, and 20 instances per class are used, respectively). While the results obtained in the supervised setting are better than those obtained in the few-shot setting, it is impressive to see that the results of the CLIP model fine-tuned with only 20 instances per class are relatively close to the results obtained in the supervised setting. Specifically, there is approximately 5% difference between the supervised CLIP results and the 20-shot results for both the informativeness and humanitarian tasks. It is also interesting to see that there is a significant jump from the results obtained with 1-shot versus 5-shot CLIP, especially for the humanitarian task. The other models also improve significantly as more instances are used and their results get close to the results of their supervised counterparts that use all the available data. Notably, CSWin has the best 1-shot result for the humanitarian task, but not for the informativeness task. These results together, and especially CLIP's few-shot results, show that a small number of instances from an emergent disaster event may be enough to obtain models that can be used to filter useful data in nearly real-time as a disaster emerges.

[RQ3] *How does CLIP perform in in-domain/cross-domain settings versus few-shot/supervised settings?* The results of the in-domain/cross-domain experiments by comparison with the few-shot/supervised experiments are shown in Table IV. As can be seen, the in-domain transfer and cross-domain transfer (all-but-one) results are generally better than the event-specific few-shot results but worse than the supervised results. However, there are some cases where these models produce better results than their supervised counterparts. Specifically, the in-domain transfer setting has better results than the baseline in 3 out of 14 cases, while the the cross-domain (all-but-one) setting has better results in 5 out of 14 case.

When analyzing the results of the cross-domain transfer (one-versus-one), we can see that they are sometimes better and sometimes worse than the few-shot results, depending on how similar the source and target events are in terms of disaster scene (e.g., Hurricane Maria/Sri Lanka Floods) or the time when they happened (e.g., Hurricane Maria/Mexico Earthquake), among others.

[RQ4] *What setting leads to the best results for an emergent disaster for which little or no labeled data is available?* While models trained/fine-tuned in the supervised setting give competitive results overall, supervised models fine-tuned with a relatively large amount of labeled data are not practical to use

for an emergent disaster as labeled data is simply not available in the early hours of the disaster. Considering the other more practical settings, the best non-supervised results for the informativeness task are split between the in-domain transfer and cross-domain transfer (all-but-one) settings, which assume no labeled data from the target disaster. However, even when the in-domain transfer results are better, the cross-domain transfer (one-but-all) models follow closely behind, making this type of model a strong candidate for filtering informative tweets in the early hours of a disaster event. Similarly, the cross-domain (all-but-one) models are also strong candidates for being used to filter tweets according to various humanitarian categories in the early hours of a disaster.

This suggests that in the early stages of a disaster, using an off-the-shelf CLIP models previously fine-tuned using as many images as available from diverse prior disasters may help achieve high performance on the target disaster. In fact, the performance may be similar to what one might obtain when using an event-specific supervised setting. As the disaster unfolds and some labeled instances from the target disaster become available, one may also consider fine-tuning a pre-trained CLIP model using directly instances from the target disaster.

VI. Error Analysis

We focus our error analysis on the humanitarian classification task. To understand what types of errors that different models make for different types of events, in Figure 1, we show confusion matrices for the best performing models, specifically, event-specific supervised models (left), event-specific 20-shot models (middle) and all-but-one models (right). From top to bottom, results are shown for a fire (D0), a hurricane (D3), an earthquake (D5) and a flood (D6). Based on the analysis of these confusion matrices, combined with analysis of correctly classified and misclassified images shown in Figure 2, several trends can be observed:

- The models often confuse the categories of *Affected* or *Effort* and *Damage*, especially when there are both humans/cars and signs of destruction in the image (for instance, images (d), (e), (m), (n), (q), (r), and (s) in Figure 2). This confusion arises because both categories can be inferred from the image. Particularly, if the destruction occupies a larger portion of the image, it is more likely to be categorized as *Damage* even if the label for the image is *Affected* or *Effort*. In some cases, multiple labels can be considered valid, as the image is related to damage, and there are people in the image affected by the disaster or making efforts to help the victims. However, each image in the dataset has only one label.
- Another issue identified, especially in the few-shot setting, is that the models have difficulty recognizing *Effort* images. These images are often categorized as *Affected* since there are people present in the image (for instance, image (l) in Figure 2). The models may need more exposure to instances where specific items such as hats, gloves, or uniforms are worn in *Effort* images to better understand their distinctive characteristics. In contrast, it is interesting to see that the few-shot models have best performance on the *Affected* class for the fire, hurricane and earthquake events.
- In *Effort* images, dogs are commonly present (for search and rescue purposes). Consequently, models sometimes misclassify images with dogs as *Effort* even if it is evident that the dogs are in distress (for instance, image (j) in Figure 2).
- Additionally, the model tends to incorrectly classify images with a significant amount of red or orange color as *Damage* since it associates these colors with fire. This association leads to misclassifications when the color is present in contexts unrelated to fire.
- While the *Damage* category has overall the largest number of false positives, it also has the largest number of true positives in all settings (according to the confusion matrices in Figure 1).
- Based on the confusion matrices, it is also interesting to see that in some cases (e.g., for the flood event), the patterns observed for supervised and all-but-one models are very similar , which shows that the diverse images in the all-but-one training subsets are sufficient to train effective models for a target event for which no labeled data is yet available.

VII. Conclusion

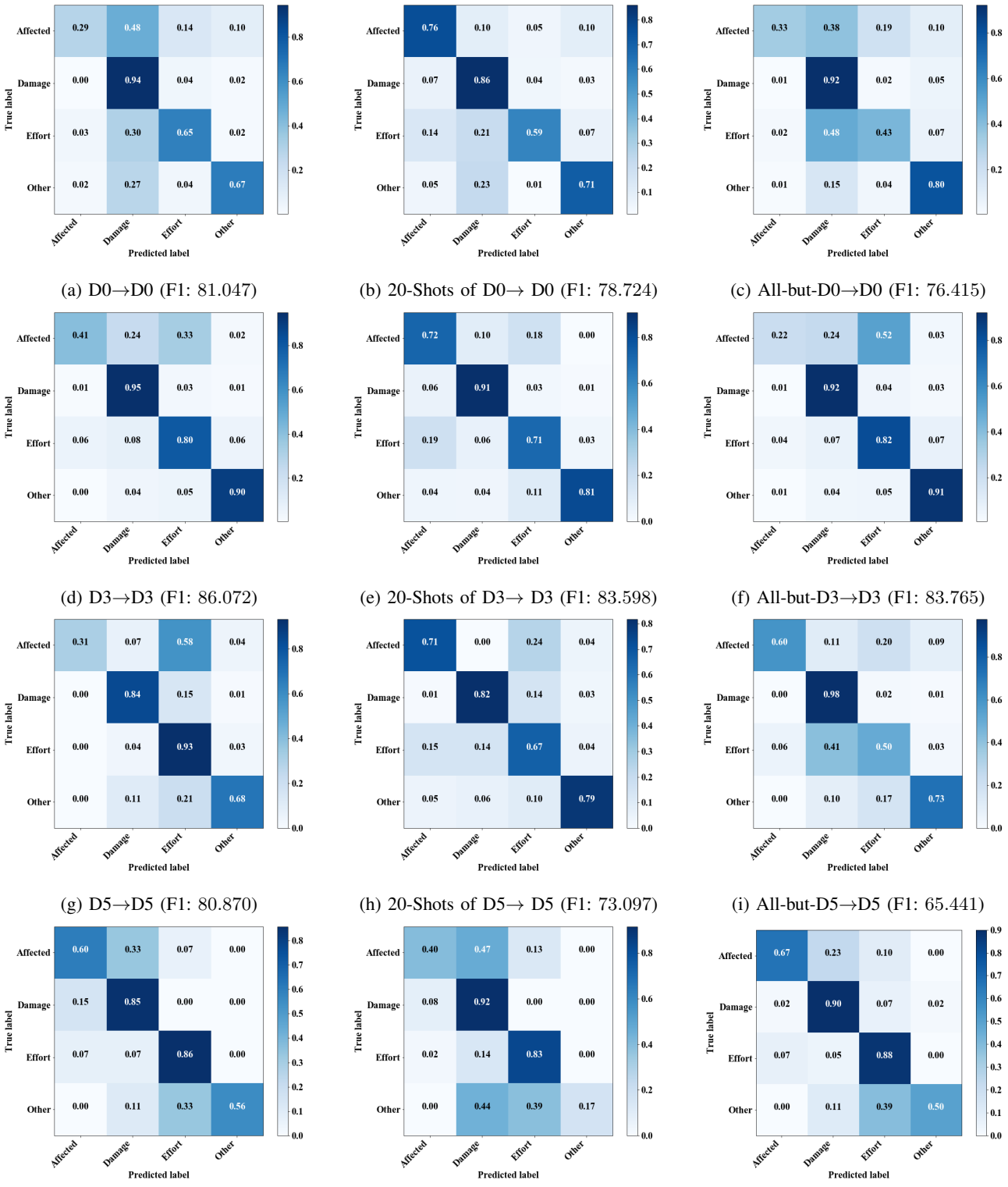
We studied the use of advanced deep learning models, specifically transformer models, ViT/CSWin, and contrastive learning, CLIP, by comparison with a CNN-based model of ConvNeXts, for disaster image classification. Experimental results showed that CLIP outperformed the transformer models and also ConvNeXts in all settings for binary-class Informativeness and multi-class Humanitarian tasks. This suggests that transformer/contrastive learning models, pre-trained on large amounts of data can capture diverse and complex patterns and can be potentially effective in disaster response efforts. Accurate and timely identification of disaster images provide useful situational awareness information and can greatly assist in response and recovery efforts. Overall, our study provides valuable insights into the potential of vision transformers and contrastive learning models for disaster response and highlights the importance of their usage in this field, to better support sustainable cities and communities and improve resilience. Future work can explore the use of other types of data, such as text, in conjunction with images to create multimodal models for disaster response efforts.

Acknowledgements

We thank the National Science Foundation and Amazon Web Services for support from grant IIS-1741345, which supported the research and the computation in this study.

References

[1] Zishan Ahmad, Raghav Jindal, NS Mukuntha, Asif Ekbal, and Pushpak Bhattachharyya. Multi-modality helps in crisis management: An attention-based deep learning approach of leveraging text for image classification. *Expert Systems with Applications*, 195:116626, 2022.



(a) D0→D0 (F1: 81.047) (b) 20-Shots of D0→ D0 (F1: 78.724) (c) All-but-D0→D0 (F1: 76.415)

(d) D3→D3 (F1: 86.072) (e) 20-Shots of D3→ D3 (F1: 83.598) (f) All-but-D3→D3 (F1: 83.765)

(g) D5→D5 (F1: 80.870) (h) 20-Shots of D5→ D5 (F1: 73.097) (i) All-but-D5→D5 (F1: 65.441)

(j) D6→D6 (F1: 76.243) (k) 20-Shots of D6→ D6 (F1: 65.997) (l) All-but-D6→D6 (F1: 79.034)

Fig. 1: The confusion matrices of Humanitarian Tasks with D0 (Fire), D3 (Hurricane), D5 (Earthquake), and D6 (Flood) as the Target Domain, across different settings of *Event-specific supervised settings*, *Event-specific 20-shot setting*, and *Cross-domain transfer settings (all-but-one)*

[2] Firoj Alam, Tanvirul Alam, Md Arid Hasan, Abul Hasnat, Muhammad Imran, and Ferda Ofli. Medic: a multi-task learning dataset for disaster

(a) (A,(A,A,A)) (b) (A,(E,A,A)) (c) (A,(D,A,D)) (d) (D,(D,E,D)) (e) (D,(E,E,E))

(f) (O,(O,O,O)) (g) (O,(O,A,O)) (h) (O,(E,E,E)) (i) (A,(A,A,A)) (j) (A,(E,E,E))

(k) (E,(E,E,E)) (l) (E,(E,A,E)) (m) (E,(E,E,D)) (n) (E,(D,D,D)) (o) (D,(D,D,D))

(p) (A,(A,A,A)) (q) (A,(A,D,A)) (r) (A,(D,D,A)) (s) (A,(D,D,D)) (t) (D,(D,D,D))

Fig. 2: Error Analysis Images for Humanitarian Tasks with D0 (Fire), D3 (Hurricane), D5 (Earthquake), and D6 (Flood) as the target event: These figures illustrate classification instances correctly classified and misclassified in various settings, including *supervised*, *20-shot*, and *all-but-one*. In each row, the left side displays the name and type of the disaster. The captions for each image follow this format: (true label, (predicted by *supervised*, predicted by *20-shots*, predicted by *all-but-one*)). True labels are depicted in black, while predicted labels are shown in green for correct classifications and in red for misclassifications. The first letter of each class represents the labels: *Others (O)*, *Affected (A)*, *Effort (E)*, and *Damage (D)*.

image classification. *Neural Computing and Applications*, 35(3):2609–2632, 2023.

[3] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Twelfth international AAAI conference on web and social media*, 2018.

[4] Firoj Alam, Ferda Ofli, Muhammad Imran, Tanvirul Alam, and Umair Qazi. Deep learning benchmarks and datasets for social media image classification for disaster response. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 151–158. IEEE, 2020.

[5] Kang An and Yanping Zhang. Lpvit: A transformer based model for pcb image classification and defect detection. *IEEE Access*, 10:42542–42553, 2022.

[6] Sourasekhar Banerjee, Yashwant Singh Patel, Pushkar Kumar, and Monowar Bhuyan. Towards post-disaster damage assessment using deep transfer learning and gan-based data augmentation. In *24th International Conference on Distributed Computing and Networking*, pages 372–377, 2023.

[7] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10231–10241, 2021.

[8] Qi Chen, Wei Wang, Kaizhu Huang, Suparna De, and Frans Coenen. Multi-modal adversarial training for crisis-related data classification on social media. In *2020 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 232–237. IEEE, 2020.

[9] De Cheng, Gerong Wang, Bo Wang, Qiang Zhang, Jungong Han, and Dingwen Zhang. Hybrid routing transformer for zero-shot learning. *Pattern Recognition*, 137:109270, 2023.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] Soudabeh Taghian Dinani and Doina Caragea. Disaster image classification using capsule networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[12] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Akash Kumar Gautam, Luv Misra, Ajit Kumar, Kush Misra, Shashwat Aggarwal, and Rajiv Ratn Shah. Multimodal analysis of disaster tweets. In *2019 IEEE Fifth international conference on multimedia big data*

(BigMM), pages 94–103. IEEE, 2019.

[15] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

[16] Hsin-Ping Huang, Deqing Sun, Yaojie Liu, Wen-Sheng Chu, Taihong Xiao, Jinwei Yuan, Hartwig Adam, and Ming-Hsuan Yang. Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In *Proceedings of the 17th European Conference on Computer Vision–ECCV 2022, Part XIII*, pages 37–54, Tel Aviv, Israel, 2022. Springer.

[17] Yusuf Selman Inanc. Turkey earthquake: Trapped victims cry for help on social media. https://www.middleeasteye.net/news/turkey-earthquake-trapped-victims\-cry-help-social-media, 2023. Accessed on 14 February 2023.

[18] Edy Irwansyah, Hansen Young, and Alexander AS Gunawan. Multi disaster building damage assessment with deep learning using satellite imagery data. *International Journal of Intelligent Systems and Applications in Engineering*, 11(1):122–131, 2023.

[19] Bo Jiang, Kangkang Zhao, and Jin Tang. Rgtransformer: Region-graph transformer for image representation and few-shot classification. *IEEE Signal Processing Letters*, 29:792–796, 2022.

[20] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[21] Larry J King. Social media use during natural disasters: An analysis of social media usage during hurricanes harvey and irma. 2018.

[22] Yilmaz Korkmaz, Salman UH Dar, Mahmut Yurt, Muzaffer Özbey, and Tolga Cukur. Unsupervised mri reconstruction via zero-shot learned adversarial transformers. *IEEE Transactions on Medical Imaging*, 41(7):1747–1763, 2022.

[23] Rani Koshy and Sivasankar Elango. Multimodal tweet classification in disaster response systems using transformer-based bidirectional attention model. *Neural Computing and Applications*, 35(2):1607–1627, 2023.

[24] Saideshwar Kotha, Smitha Haridasan, Ajita Rattani, Aaron Bowen, Glyn Rimmington, and Atri Dutta. Multimodal combination of text and image tweets for disaster response assessment. International Workshop on Data-driven Resilience Research, 2022.

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

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[27] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[28] Zhihao Ma, Wei Li, Muyang Zhang, Weiliang Meng, Shibiao Xu, and Xiaopeng Zhang. Htcvit: an effective network for image classification and segmentation based on natural disaster datasets. *The Visual Computer*, pages 1–13, 2023.

[29] Zhihao Ma, Mengke Yuan, Jiaming Gu, Weiliang Meng, Shibiao Xu, and Xiaopeng Zhang. Triple-strip attention mechanism-based natural disaster images classification and segmentation. *The Visual Computer*, 38(9-10):3163–3173, 2022.

[30] Sreenivasulu Madichetty. Classifying informative and non-informative tweets from the twitter by adapting image features during disaster. *Multimedia Tools and Applications*, 79:28901–28923, 2020.

[31] Sreenivasulu Madichetty, Sridevi Muthukumarasamy, and P Jayadev. Multi-modal classification of twitter data during disasters for humanitarian response. *Journal of ambient intelligence and humanized computing*, 12:10223–10237, 2021.

[32] Sayak Nag, Orpaz Goldstein, and Amit K Roy-Chowdhury. Semantics guided contrastive learning of transformers for zero-shot temporal activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6243–6253, 2023.

[33] Ganesh Nalluru, Rahul Pandey, and Hemant Purohit. Relevancy classification of multimodal social media streams for emergency services. In *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 121–125. IEEE, 2019.

[34] Ferda Ofli, Firoj Alam, and Muhammad Imran. Analysis of social media data using multimodal deep learning for disaster response. *arXiv preprint arXiv:2004.11838*, 2020.

[35] Yishu Peng, Yaru Liu, Bing Tu, and Yuwen Zhang. Convolutional transformer-based few-shot learning for cross-domain hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[37] Pradeep Kumar Roy, Abhinav Kumar, Jyoti Prakash Singh, Yogesh Kumar Dwivedi, Nripendra Pratap Rana, and Ramakrishnan Raman. Disaster related social media content processing for sustainable cities. *Sustainable Cities and Society*, 75:103363, 2021.

[38] Linus Scheibenreif, Joëlle Hanna, Michael Mommert, and Damian Borth. Self-supervised vision transformers for land-cover segmentation and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1422–1431, 2022.

[39] Iustin Sirbu, Tiberiu Sosea, Cornelia Caragea, Doina Caragea, and Traian Rebedea. Multimodal semi-supervised learning for disaster tweet classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2711–2723, 2022.

[40] Maryam Sultana, Muzammal Naseer, Muhammad Haris Khan, Salman Khan, and Fahad Shahbaz Khan. Self-distilled vision transformer for domain generalization. In *Proceedings of the Asian Conference on Computer Vision*, pages 3068–3085, 2022.

[41] Tao Sun, Cheng Lu, Tianshuo Zhang, and Haibin Ling. Safe self-refinement for transformer-based domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7191–7200, 2022.

[42] Congcong Wang, Paul Nulty, and David Lillis. Crisis domain adaptation using sequence-to-sequence transformers. *arXiv preprint arXiv:2110.08015*, 2021.

[43] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

[44] Xiyu Wang, Pengxin Guo, and Yu Zhang. Domain adaptation via bidirectional cross-attention transformer. *arXiv preprint arXiv:2201.05887*, 2022.

[45] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[46] Chengming Xu, Siqian Yang, Yabiao Wang, Zhanxiong Wang, Yanwei Fu, and Xiangyang Xue. Exploring efficient few-shot adaptation for vision transformers. *arXiv preprint arXiv:2301.02419*, 2023.

[47] Huali Xu, Shuaifeng Zhi, Shuzhou Sun, Vishal M Patel, and Li Liu. Deep learning for cross-domain few-shot visual recognition: A survey. *arXiv preprint arXiv:2303.08557*, 2023.

[48] Li Xukun and Doina Caragea. Improving disaster-related tweet classification with a multimodal approach. In *ISCRAM 2020 Conference Proceedings–17th International Conference on Information Systems for Crisis Response and Management*, 2020.

[49] Jinyu Yang, Jingjing Liu, Ning Xu, and Junzhou Huang. Tvt: Transferable vision transformer for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 520–530, 2023.

[50] Chen Zhao, Renjun Shuai, Li Ma, Wenjia Liu, and Menglin Wu. Improving cervical cancer classification with imbalanced datasets combining taming transformers with t2t-vit. *Multimedia tools and applications*, 81(17):24265–24300, 2022.