IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 June 2024, accepted 7 August 2024, date of publication 13 August 2024, date of current version 11 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3443130

RESEARCH ARTICLE

Identifying Disaster Regions in Images Through Attention Shifting With a Retarget Network

NARONGTHAT THANYAWET[1], (Member, IEEE), PHOTCHARA RATSAMEE[2], (Member, IEEE), YUKI URANISHI[1], (Member, IEEE), MASATO KOBAYASHI[1], (Member, IEEE), AND HARUO TAKEMURA[1], (Life Member, IEEE)

[1]Graduate School of Information Science and Technology, Osaka University, Suita, Osaka 565-0871, Japan
[2]Graduate School and Faculty of Robotics and Design, Osaka Institute of Technology, Osaka 535-8585, Japan

Corresponding author: Narongthat Thanyawet (narongthat.thanyawet@lab.ime.cmc.osaka-u.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Scientific Research (B) under Grant 22H01449, in part by the KAKENHI Fund for the Promotion of Joint International Research [Fostering Joint International Research (B)] under Grant JP20KK0086, and in part by the Mohamed Bin Zayed International Robotics Challenge (MBZIRC) Grant.

ABSTRACT Disasters disrupt lives and necessitate quick location of affected areas for rescue efforts. The application of computer vision has enhanced disaster detection, such as landslides and floods; however traditional computer vision methods often overlook smaller, critical details in favor of prominent objects. This research introduces the Retarget Network (RetNet), a novel framework aimed at improving image captioning techniques to identify and prioritize these less evident, yet crucial, objects in disaster scenarios, enhancing scene recognition and aiding in more effective disaster response. By masking images, we direct the model's focus towards additional significant areas within the image. RetNet employs anchor boxes to refine the targeting of specific areas and, optimize their center positions, heights, and widths. Additionally, RetNet determines which anchors to mask prior to captioning, facilitating the identification of challenging objects such as boulders, soil, and water, which are indicative of natural disasters. We validated RetNet across multiple disaster scenarios—landslides, floods, and wildfires—using images taken from various perspectives, including side-view, aerial and shipborne views. Our findings reveal an accuracy of 91.60% in landslide detection from side-view image captions and 87.50% for detections from shipborne views. These results underscore RetNet's effectiveness in enhancing the identification of disaster-affefted regions.

INDEX TERMS Retarget network, image captioning, vision transformer, anchor box, disaster detection.

I. INTRODUCTION

Disasters such as landslides [1], flooding [2], [3] and wildfires [4], can occur anywhere and significantly affect people's lives and daily activities. Landslides, particularly prevalent in mountainous regions, are triggered by various factors, including intense rainfall, slope instability, or seismic events. When disasters such as landslides occur, they can block transportation routes and damage critical infrastructure, such as roads, buildings, and highways. It is crucial for decision-makers to accurately identify affected areas to plan effectively for reconstruction and recovery. Unmanned Aerial Vehicles (UAVs) play a vital role in this process, providing valuable image data to assess damage and guide decision-making. To overcome these issues, numerous computer vision approaches, including detection and classification, have been developed to improve the effectiveness and speed of disaster detection.

Traditional image classification challenges typically involve identifying a single class for each image, as noted in various studies [5], [6]. Many techniques in this domain leverage neural networks to extract image features, thereby facilitating the classification process [7], [8], [9], [10], [11], [12], [13]. Following advancements in image classification, object detection emerged as a pivotal technique, with the aim of identifying more specific classes within each boundary

The associate editor coordinating the review of this manuscript and approving it for publication was Joao Neves.

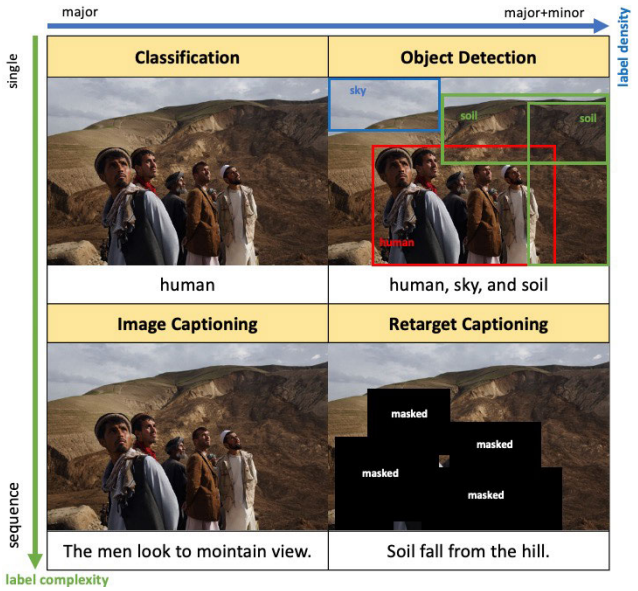


FIGURE 1. Concentrated captioning for more focusing and explanation on the target objective.

in an image. Anchor boxes are used to define bounding boxes [14], [15], [16] for target object classes in object detection. Despite these advancements, the identification of landslides, floods, and wildfires remains challenging. This difficulty often stems from the frequently blurry appearance of affected regions, such as landslide areas resembling typical soil or flooded areas mirroring the appearance of ponds or lakes [2], [3], leading to potential misclassification.

However, the emergence of image captioning has brought to light a more nuanced dimension of computer vision. This goes beyond simple classification to embrace the description of an image's surrounding context. Traditionally, end-to-end methods for image captioning involved feature extraction from images using Convolutional Neural Networks (CNNs), followed by the generation of textual captions using Recurrent Neural Networks (RNNs) [17]. More recently, the introduction of transformers [18] has revolutionized this field. This approach encodes an image into a series of image tokens, and the textual caption is similarly decoded into text tokens. These tokens are then processed through an encoder-decoder attention layer, known as Vision Transformers (ViT) [19], to establish relationships between the image and the caption content. Even though current image captioning models can generate detailed captions of images, thereby enhancing disaster detection, as evidenced in existing research [20], these models still primarily focus on major objects to generate captions, as depicted in Figure 1. This approach often overlooks smaller yet significant regions within an image. In particular, aerial images that capture expansive areas may contain small regions where disasters exist. Developing a model that recognizes not only the main objects in a scene but also generates captions describing other important parts of the surroundings in the image could enable quicker emergency response.

In this study, we propose a retargeting network (RetNet) that directs the model's focus toward specific target objectives, even though it is a small region of the disaster in the images. Drawing inspiration from [17], we utilize anchor boxes and optimize their central positioning, height, and width to define the initial masked areas. Furthermore, RetNet includes a secondary objective: selecting the most appropriate anchor boxes for masking that are aligned with the targeted objectives. This novel approach enables the model's focus on significant objects that offers a more detailed and contextual understanding.

Our contributions in this field are threefold:

- This study introduces RetNet, a groundbreaking approach that reorients the attention of image captioning models. Traditionally, these models prioritize larger and more prominent objects in an image. However, RetNet is designed to highlight more minor yet significant elements, a feature that is particularly beneficial in complex natural environments. These minor elements often contain critical information that conventional methods can overlook in such settings.
- Building upon the foundations of the Fast-RCNN framework, this study demonstrates an advanced technique for optimizing anchor boxes. This method focuses explicitly on fine-tuning the anchor boxes's central positioning, height, and width. Such optimization is crucial for improving the detection and subsequent captioning of smaller objects within natural scenes, a task that standard image captioning models may find challenging.
- The study broadens the validation of the RetNet model across a diverse range of disasters, including floods and wildfires, and examines these from various perspectives such as aerial, shipborne, and conventional human-captured views. RetNet was employed to investigate areas affected by disasters, and the model's enhanced image captioning and text classification capabilities were leveraged to accurately identify and categorize various aspects and types of disasters. This application demonstrates the practical and significant impact of RetNet in real-world scenarios, particularly in term of timely and effective disaster response and assessment.

II. RELATED WORKS

This section reviews related work on the evolution of computer vision techniques applicable to disaster detection, including traditional image classification [5], [6], [7], [8], [10], [11], object detection [13], [21], object segmentation [9], [12] and image captioning generation [17], [22]. Despite these advanced techniques, accurately identifying hazy or unclear objects in images, such as during disasters, remains a challenge.

A. ADVANCEMENT IN OBJECT DETECTION

The use of anchor boxes to identify objects in images is a cutting-edge computer vision technique. This method eliminates low-probability detection during inference and

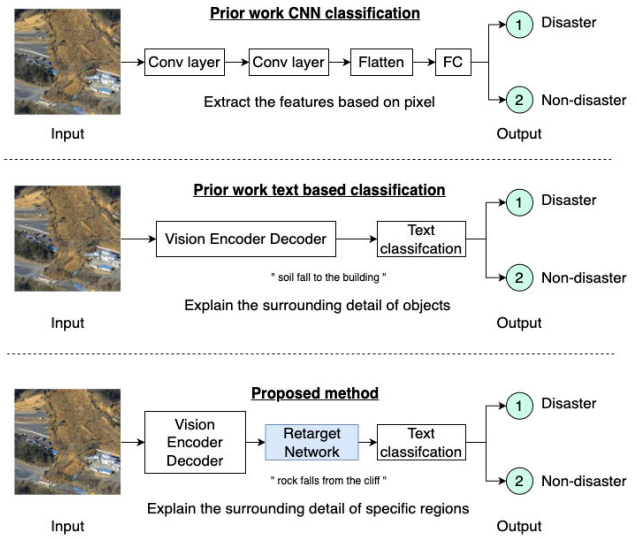


FIGURE 2. Disaster image classification network architecture between prior works and our proposed method. There are convolution layers and flatten layer for prior work. For proposed method, VED to generate caption token with masked images from Retarget Network for text classification.

creates anchor boxes across each image patch. In order to preserve only the highest probability detection, the procedure uses intersection over union (IoU) computations for every class, as explained in important studies such as R-CNN, Fast R-CNN, and Faster R-CNN [14], [15], [16]. In the training phase, labeling the object classes and boundary boxes is an essential component of this technique. Similar to image segmentation, where boundary labels are used for feature extraction prior to training, this preparation is necessary for the model to be trained to produce anchor boxes accurately.

Expanding upon the concept of anchor boxes, our methodology introduces novelty by re-purposing the anchor box generation process to generate masked regions within the image. The ideal locations and dimensions of these anchor boxes were taught to the model during the training phase, after which they were employed to mask the original image. This method aims to refocus the model's attention on less accurate and, more frequently missed objects in an image. We wanted to improve the model's capacity to identify and determine uncertain or less notable elements in the visual data by changing the attentions hierarchy of the model.

B. TRANSFORMER-BASED IMAGE CAPTIONING

In recent years, transformer models have marked a significant advancement in the field. Initially developed for machine translation, these models utilize attention layers comprising of transformer blocks. These blocks significantly enhance the model's focus, allowing it to concentrate more effectively on values that exhibit relevant relationships. This development represents a substantial shift in the processing and interpretation of visual data is processed and interpreted, offering a more nuanced and contextually aware approach to computer vision tasks. Following their success in natural language processing, transformer models have been adapted for computer vision, particularly in image classification [19], [23], [24]. Applying transformer techniques in this context enhances feature extraction, allowing for more focused attention on significant features within images. This development marks a notable evolution in analyzing visual information, enabling more accurate and contextually rich interpretations.

In the text captioning part, tokenizers are used for embedding the text into vectors. Several tokenizers are used in image captioning architectures, such as BERT [25] based on the WordPiece technique, DistilBERT [26], a smaller pretrained model derived from BERT, and GPT-2 [27] based on the Byte Pair Encoding (BPE) technique.

C. CHALLENGE IN CURRENT IMAGE CAPTIONING

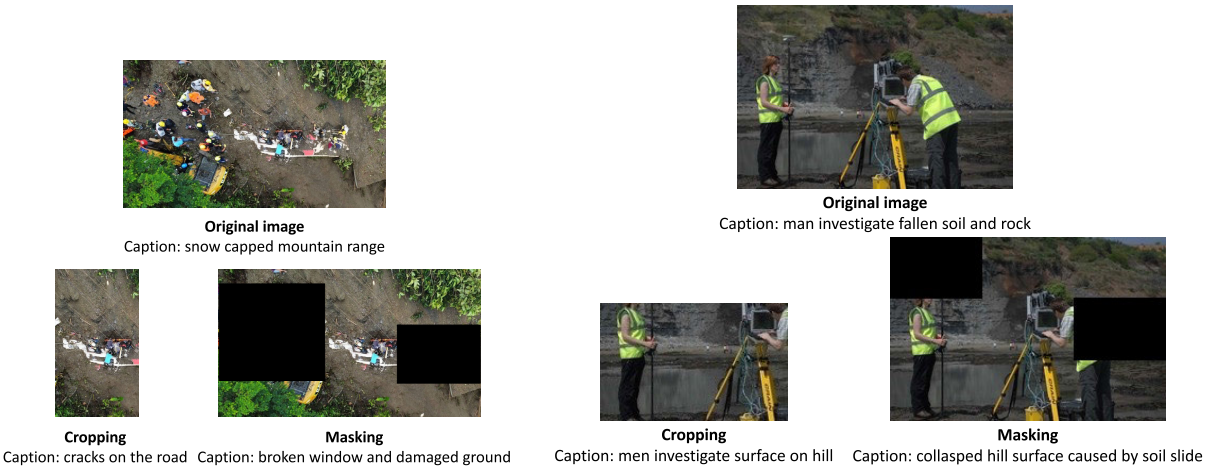
A critical area of computer vision research is the production of text captions from input images. RNNs are typically used for caption generation, after CNNs for feature extraction [17], [22]. This method efficiently creates a cohesive narrative for the visual data by connecting the text captioning with the extracted image features. Additionally, thorough captioning has been used in light of the intricacy of the images, which frequently feature numerous objects. Using anchor boxes, this technique crops the image into regions of interest and creates a unique caption for each region, providing a more thorough explanation [17].

We propose that an image should contain several elements, each of which deserves a more thorough explanation than a single sentence. Image captioning provides a more meaningful way to express these complexities, especially when generating outputs in human-language terms. This method works particularly well for elucidating ambiguous objects, such as those discovered during natural disasters. Object features only represented as pixel-based images may need to be precisely identified or categorized. As a result, image captioning can play an essential part in giving people with a deeper understanding of these complicated scenes.

D. APPLICATION IN DISASTER MANAGEMENT

According to the 2P2R approach [28], disaster management starts with preventative actions such as building or remodeling infrastructure to reduce the effects of disasters. Building sea walls to prevent tidal [29] and redesigning building foundations to absorb seismic vibrations more effectively [30], [31] are two examples. The next step is preparation, which includes pre-event plans like securing food and water supplies and mapping out evacuation routes. After an incident, the recovery phase evaluates and repairs the damage, including identifying victims and damaged areas. The response phase entails warning and evacuation people. UAV and helicopter aerial reconnaissance is invaluable during this phase.

In order to define areas that require restoration, our research attempts to apply our novel approach during the recovery phase using images to identify damaged regions.



(a) Aerial view of landslide image.

(b) Side view of landslide image.

FIGURE 3. Landslide image captioning between applying cropping and masking techniques to focus on the affected area in different view aspect.

TABLE 1. Comparison of existing methods for image classification.

Method	Accuracy	Precision	Recall	F1-Score
ResNet50 fine-tune [10]	67.58	67.58	**100.00**	80.65
ResNet50 from scratch [10]	71.61	**98.81**	70.77	82.47
VED [20]	**95.00**	96.19	96.42	**96.31**

Natural features are frequently displaced into unusual locations during disasters, making it challenging to identify these uncertain objects with conventional methods. Determining whether soil sliding from a cliff and obstructing a mountain transportation route constitutes a damaged region is difficult. Through the improvement of such unclear object detection, our method has the potential to significantly increase the precision and effectiveness of damage assessment in disaster management.

III. PRELIMINARY INVESTIGATION

In this section, we explore caption-based [20], which demonstrate that the classification challenge outperforms pixel-based techniques [10]. The caption-based model can achieve more meaningful features, making the model better classified in disaster images. Moreover, the caption-generated machine-learning models [17], [18] tend to focus on most objects in the foreground. This experiment demonstrates the model's attention to obtain valuable features from masking and cropping images and the characteristics of optimal masking to achieve the target objects.

A. CLASSIFICATION (PIXEL-BASED VS CAPTION-BASED)

1) EXPERIMENT SETUP

Our goal in this experiment was to demonstrate that traditional pixel-based models, which rely on feature extraction for image classification, face significant challenges when dealing with disaster scenes. We compare disaster scene classification between pixel-based models (specifically, ResNet50 [10]) and text-based image captioning features that utilize a Vision Encoder-Decoder (VED) framework. We used 620 images as the testing dataset, trained on 5,280 images, and set aside 605 images for validation. The image dataset contains both normal and landslide scenes.

2) EXPERIMENTAL RESULTS

The results are presented in Table. 1. Using 620 images in the testing dataset demonstrated that the VED approach significantly outperformed traditional models, achieving an Area Under the Curve (AUC) of 0.94 [20], which is a notable improvement over the benchmark set by ResNet50 [10]. However, we noticed that the traditional ResNet50 models and the VED method tend to prioritize foreground objects, often missing small disaster areas in the image. This observation inspired us to develop a framework to shift the focus toward surrounding objects, such as small disaster regions, as represented in Figure 2. The Retarget Network adjusts the prioritized regions to better detect target areas in disaster images, particularly large-scale images.

B. IMAGE CAPTIONING FROM CROPPING VS MASKING

1) EXPERIMENT SETUP

We experimented with cropping and masking as two distinct image-editing techniques. Cropping entails reducing the image's outer regions to concentrate on particular sections, which may cause the model to pay more attention to typically overlooked items. Masking is the process of hiding portions of an image so that the model can focus on the exposed sections, which may be less noticeable but are essential. We utilized 16 represented images from the side view, shipborne, and aerial view images in these experiments.

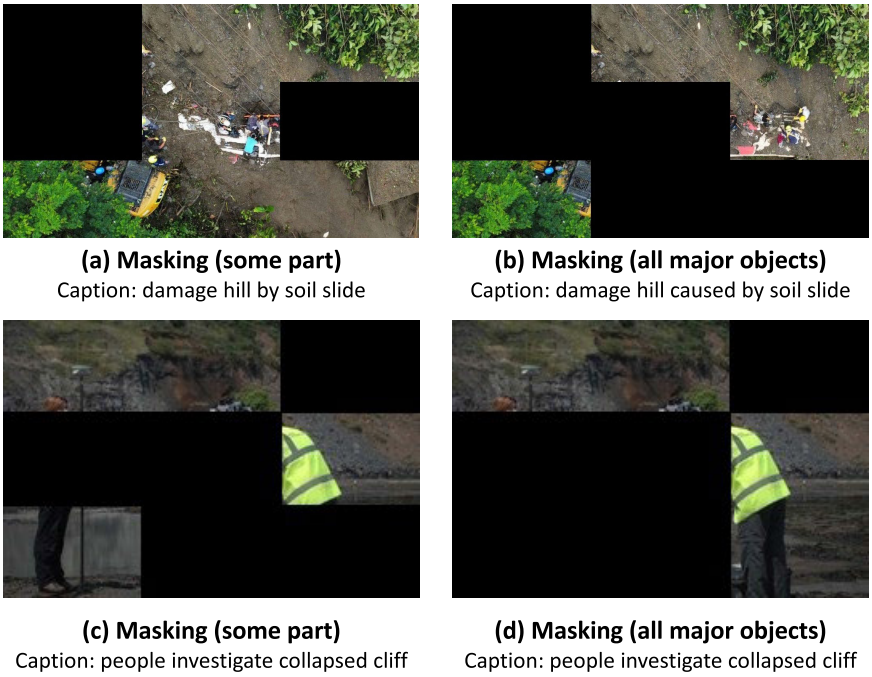


FIGURE 4. Comparison of images with only some parts masked ((a) and (c)) and with all major objects masked ((b) and (d)).

2) EXPERIMENTAL RESULTS

The analysis (as shown in Figure 3) reveals differing results from the two experimental methods. One significant drawback of cropping images was the reduction in significant features from surrounding objects. Due to this scarcity, captions were often misleading or failed to provide sufficient information for a comprehensive explanation in common language.

However, the masking method yielded more encouraging results. The model produces captions that are included these secondary objects owing to masking, which selectively covered the main focus object while leaving other components visible. This method allows for a more comprehensive understanding of the scene because the captions can explain one specific object, the larger context and the relationships between the various components. In particular, masking can direct the model's focus toward a more impartial and thorough understanding of the disaster scene.

C. OPTIMAL MASKING

1) EXPERIMENT SETUP

Similar to the cropping vs masking experiment, we also utilized 16 represented images from the side view, shipborne, and aerial view images in these experiments. We then examined how masking particular areas affects a model's capacity to produce accurate image captions. Using a brute-force process to cover every potential region, we determined the characteristics of the covered areas that produced captions closest to the actual data. In order to mask each image for this experiment, we divided it into nine patches, which were arranged in a 3×3 grid. As a result, 512 distinct masking configurations could be made, or $2^{(3\times3)}$. We examined the characteristics and trends in the captions created for the masked images for all possible combinations.

2) EXPERIMENTAL RESULTS

Our results demonstrate that accurate caption generation does not require masking every noticeable objects in the image. Figure 4 represents this example, illustrating how the captions of images with some masking and images with all masking major objects closely resemble each other. Furthermore, we observed that dealing with 16 images using the brute-force masking method required more than 4 hours of computing time. These results highlight that masking significantly raises computational demands, even though it can effectively shift the model's attention. This emphasizes the requirement for enhanced approaches. As a result, we introduce the Retarget Network, a novel method designed to balance computational efficiency and high-quality detection.

IV. METHODOLOGY

In this section, we introduce the network architecture and pipeline of our proposed RetNet, designed to shift attention from major object to others, which may be less noticeable but are essential parts of the image.

A. NETWORK ARCHITECTURE

The RetNet's architecture is illustrated in Figure 5. Before extracting the image features into feature maps using the

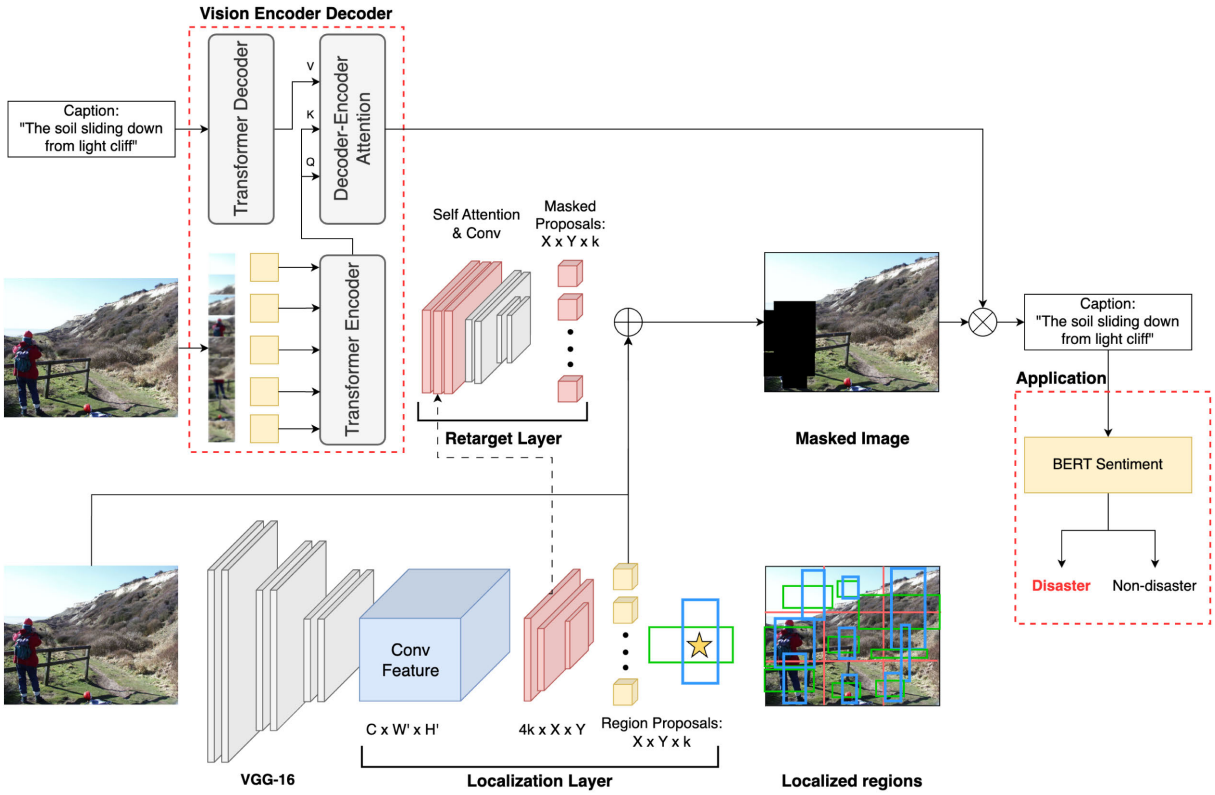


FIGURE 5. Our network architecture.

VGG-16 model, we trained the VED for image captioning [17]. We add two novel layers to the RetNet architecture in this network: the Localization layer and the Retarget layer. The retarget layer then refines the possible masked region candidates produced by the localization layer to optimal masks. In the localization layer, patches and anchor boxes were used to create these candidate-masked regions. The focus of the image captioning model is then shifted using the retarget layer to recognize the most significant masked regions among the candidates.

B. VISION ENCODER DECODER

Image captioning using a Transformer model in a VED [19], [23], [24] framework generates a caption for an image by combining computer vision and natural language processing techniques. The two main components of the framework were the text decoder and the vision encoder.

- Vision Encoder: An image is divided into patches and, then fed into encoder transformers block to feature vectors. In this process, we used a pre-trained from ViT model [19] to encode the images.
- Caption Decoder: The text caption label is embedded into feature vectors using the decoder transformer block [18] from the input text. For the decoder process, we used the Bidirectional Encoder Representations from Transformers (BERT) [25] embedding the caption text including tokenizers. The BERT tokenizer, based on the WordPiece technique, is suitable for use with natural objects in environmental scenes in our challenge.

We utilize a dataset of image pairs at fixed-length captions to train the network to learn the correlations between the feature vectors and their corresponding textual descriptions [32]. In order to generate text captions based on the correlations identified in the image feature vectors, the model applies learned parameters during the inference process.

C. LOCALIZATION LAYER

This layer aims to generate candidates for masking regions from the feature map [33] of an image. We employ the state-of-the-art VGG-16 architecture [34], [35], which consists of 13 layers of 3×3 convolutions alternated with 5 layers of 2×2 max pooling, to extract the feature map I. The size of the input image, width W, and height H that we transform for the pipeline into 512. As a result, an input image with dimensions $3 \times W \times H$ is transformed into a feature map with dimension $C \times W' \times H'$, where $C = 512$, $W' = \lfloor \frac{W}{16} \rfloor$, and $H' = \lfloor \frac{H}{16} \rfloor$.

We modified this technique to generate candidates for masking regions, illustrating inspiration from the Region Proposal [17], that uses patches and anchor boxes to generate dense captions. We first divided the images into square grid patches containing k anchor boxes to cover almost all probable objects. We use feature maps into 7×7 grid patches to ensure the coverage of even small objects in

the image. Four parameters are specified for each patch: its width w_p, height h_p, and center position (x_p, y_p). Using a regressive offset model represented by the following equations, we define k anchor boxes within each patch with four parameters for size (w_a, h_a) and the center position (x_a, y_a) of each anchor box's region:

$$x_a = x_p + t_x \frac{w_p}{2} \tag{1}$$

$$y_a = y_p + t_y \frac{h_p}{2} \tag{2}$$

$$w_a = w_p \cdot \exp(t_w) \tag{3}$$

$$h_a = h_p \cdot \exp(t_h) \tag{4}$$

In this instance, the normalized offset from the anchor's center is represented by (t_x, t_y) which uses the hyperbolic tangent activation function (Tanh) in the range $(-1, 1)$ to ensure that the center position of the anchor boxes is in the patch size, and the log-scale transformation of the anchor size is represented by $t_w, t_h)$ which used the rectified linear unit (ReLU) in the range $[0, \infty)$ to ensure that the width and height of the anchor boxes could cover the objects. We then map (x_a, y_a, w_a, h_a) onto the input feature map's $X \times Y$ grid and then transform the feature map back to its original dimensions of $W \times H$.

D. RETARGET LAYER

Along with the input from the feature map for the self-attention process, we obtain a set of anchors $R(x_a, y_a, w_a, h_a)$ from the localization layer for use as candidates. $I \in \mathbb{R}^{B\times C\times W\times H}$ is the feature map that we used, where B is batch size, C for number of channels which is 512, W for width, and H for height. The attention map A is first calculated as follows:

$$A(Q, K) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{5}$$

In this section $Q, K \in \mathbb{R}^{B\times \frac{C}{8}\times N}$ denotes the query and key matrices retrieved from the feature map. The scaling factor for the dimensionality of the critical vectors, in our case $n = 8$, is denoted by the term $d_k = \frac{C}{n}$ [18].

Subsequently, we utilize the attention map A to a single dimension of $N = W \times H$ using the value of the feature map matrix V^T. The output feature map is obtained by multiplyingattention map A by V. The output of masking region O is then obtained by scaling this result by a learnable parameter γ and adding the input feature map I, which includes a residual connection as follows:

$$O(A, V, I) = \gamma \cdot Reshape(AV^T) + I \tag{6}$$

After that, the output is reshaped using the Reshape() function to return it to the original spatial dimensions, $W \times H$. Then, O is then subjected to the sigmoid function to determine the optimal masking anchor boxes M as follows:

$$M = Sigmoid(O) \odot R \tag{7}$$

In this particular instance, the dimensions of the anchor box candidates R are $X \times Y \times 4k$. Masking region from anchor box candidates M calculated from the result Sigmoid(O) and anchor box candidates R with the Hadamard product from these two metrics. As a kind of optimal masking, we later mapped M with dimensions of $X \times Y \times k$ back to the original images $X \times Y$ dimension images. Maintain consideration that, for each anchor parameter, we mask the original images using a masking value M of 1, denoted by Sigmoid$(O) > 0.6$. We identify non-masking areas with a masking value of 0, denoted by Sigmoid$(O) <= 0.6$. Then, to create captions C_{gen}, we fed the masked images into the VED framework [19], [23], [24].

E. LOSS FUNCTION

During the training phase, we used each image's reference captions C_{true} as the ground truth. We employed a smooth L1 regularization $L1_1^{reg}$ in the transformed coordinate space [36] to penalize the parameters to compare the generated captions with the ground truth captions. In order to measure the similarity between the captions, we additionally employed the inverse cosine similarity loss L_{invc}.

An encoding and embedding process is necessary to calculate loss. Utilizing a model known for its efficacy in generating contextual embeddings, BERT [25], we generated captions C_{gen} based on the reference captions C_{true}. We addressed both C_{gen} and C_{true} using the BERT tokenizer to generate the encoded vectors. After that, we produce the embeddings V_{gen} and V_{true} from the mean of the encoded vectors' final hidden states as follows:

$$V_{gen} = Mean(BERT(Encode(C_{gen}))) \tag{8}$$

$$V_{true} = Mean(BERT(Encode(C_{true}))) \tag{9}$$

After obtaining the embeddings V_{gen} and V_{true}, we used them to calculate the inverse cosine similarity loss, L_{invc}, as follows:

$$L_{invc} = \frac{V_{gen} \cdot V_{true}}{\|V_{gen}\|\|V_{true}\|} \tag{10}$$

L1 regularization penalizes the model parameters to prevent over-fitting. It applies the absolute values of the parameters, which do not impose excessive penalties on the loss function. It generates sparse solutions for feature selection from the parameters in the RetNet layers, which contain quite large features. We also calculated the smooth L1 regularization L_1^{reg} in the transform coordinate space as follow:

$$L_1^{reg} = \sum_{p\in Parameters} \|p\|_1 \tag{11}$$

Lastly, we use a custom loss function in our model training phase, which combines the cosine similarity loss L_{invc} and the smooth L1 regularization L_1^{reg}, each of which has weights β and α assigned to it, respectively. The following is the combined loss equation:

$$L_{custom} = \alpha \cdot L_{invc} + \beta \cdot L_1^{reg} \tag{12}$$

TABLE 2. The statistics of our dataset.

Data source	Training	Validation	Testing	Total	Type
BGS [37]	1,690	200	146	2,036	Disaster
Normal [38]	3,500	669	150	4,319	Normal
DID [39]	550	70	114	734	Disaster
Shipborne [40]	-	-	270	270	Diaster and Normal
Total	5,740	939	680	**7,359**	

The computational complexity of the RetNet architecture in training process as the training process: $O\left(\frac{n}{b}\right) + O\left((W \cdot H)^2 \cdot C\right) + O(p)$, in the inference process: $O\left((W \cdot H)^2 \cdot C\right)$

V. EXPERIMENTS AND RESULTS

Our method involves the processing of images using the *RetNet* to produce region proposals that are fed into RetNet. The outcomes of our suggested approach are shown in this section, along with how it performed with various ablation loss functions, patch grid counts, and anchor box counts. Moreover, we use our network to analyze other disasters, floods, and wildfires and generate captions for the various scenarios. Furthermore, we utilize our system to analyze landslide scenarios from an alternative viewpoint (i.e., shipborne imagery) and assess its efficacy compared to conventional detection approaches.

A. DATASET

In this study, we utilized image datasets from four primary sources, as detailed in Table 2:

- The British Geological Survey (BGS) [37] provides landslide images. We further annotated these images using text captions to enhance our dataset for the intended analyses.
- Kaggle [38] contributed an extensive collection of 4,319 images, from which we derived a common scene image dataset.
- The Disaster Image Dataset (DID) [39] provided the flood and wildfire images. We annotated the images with captions in our network.
- Shipborne [40] provided landslide and non-landslide imagescaptured during a survey conducted on a ship.

The models in our framework were trained, validated, and tested using BGS, Kaggle, and DID data sources. The Shipborne dataset was only used for testing purposes in classification application. According to Table 2, there are 2,036 abnormal images from BGS and 4,319 typical images from Kaggle. Additionally, we extended the disaster datasets for additional disasters consist of floods and wildfires, as the DID [39]. We decided to include these two beyond disaster segments in our studies to explore the various scenarios.

We made an effort to use label terms frequently found in this dataset, such as trees, rocks, dirt, water, rivers, fires, and lakes. A text caption was applied to the dataset. We also described the surrounding objects and their locations using the terms of the object since this would make it simpler for the model to understand the conditions of the scenes.

We used the shipborne image-based landslide dataset [40], which includes 270 images, 231 of which have no relation to landslides and 39 of which do. As a result, we could use the classification on an alternative dataset. This dataset serves as one of the benchmarks we use to evaluate the effectiveness of our proposed approach.

B. ABLATION STUDY

Our approach relies on image masking to retarget the attention during image captioning. Girshick [15]'s recommended an L1 regularization function for parameters, such as the anchor box form. In order to optimize the retarget layer and provide captions comparable to the labels, we employed a cosine similarity loss function. We also constructed an inverse cosine similarity loss function to find dissimilar captions. This method encourages the model to mask images to highlight target objects, which is novel for detecting unique objects through an inverse function. Images and captions from our collection, such as "The man in front of the rocky mountain, whose soil has collapsed," feature a variety of subjects outside of disaster-related themes.

The model performance is significantly impacted by the L1 regularization function, as indicated by the findings presented in Table 3. BLEU [41], METEOR [42], ROUGE-L [43], and CIDEr [44] are the standard linguistic metrics we utilize to assess the relevance of text captions. The results showed that the cosine similarity loss function is not as effective as the inverse cosine similarity loss function. The most effective set of loss functions, inverse cosine similarity and L1 regularization, yielded scores of 0.416, 0.339, 0.298, 0.258, 0.270, 0.401, and 1.799 for BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L, and CIDEr, respectively. Even though the inverse cosine similarity slightly outperforms the cosine similarity loss, adding an extra loss function does not enhance the results beyond the combination of inverse cosine similarity and L1 regularization.

Furthermore, as indicated in Table 4, we tested with 3, 4, 5, and 7 patch grids, varying the setup by employing three anchors in each grid. In addition, we experimented with seven patch grids utilizing three, five, and seven anchor boxes to find configurations that maximize the model performance,

TABLE 3. The ablation of loss functions in our proposed method.

Cosine	Inv Cosine	L1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
✓	–	–	0.360	0.276	0.234	0.199	0.224	0.350	1.260
–	✓	–	0.378	0.295	0.253	0.218	0.237	0.362	1.470
–	–	✓	0.415	**0.339**	0.297	**0.258**	**0.270**	0.400	1.770
✓	–	✓	0.415	**0.339**	0.297	**0.258**	**0.270**	0.400	1.770
–	✓	✓	**0.416**	**0.339**	**0.298**	**0.258**	**0.270**	**0.401**	**1.779**
✓	✓	✓	**0.416**	**0.339**	**0.298**	**0.258**	**0.270**	**0.401**	**1.779**

TABLE 4. The ablation of number of patch grid with 3 anchor boxes in our proposed method.

Number of patch grid	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
3	0.416	0.339	0.298	0.258	0.270	0.401	1.779
4	0.416	0.339	0.298	0.258	0.270	0.401	1.779
5	0.416	0.339	0.298	0.258	0.270	0.401	1.779
7	0.416	0.339	0.298	0.258	0.270	0.401	1.779

TABLE 5. The ablation of number of anchor boxes with 7 patch grids in our proposed method.

Number of anchor boxes	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
3	0.416	0.339	0.298	0.258	0.270	0.401	1.779
5	0.416	0.339	0.298	0.258	0.270	0.401	1.779
7	0.416	0.339	0.298	0.258	0.270	0.401	1.779

TABLE 6. Performance metrics for landslide, flood, and wildfire disaster in image captioning.

Disaster	BLEU-1	METEOR	ROUGE_L	CIDEr
Landslide	0.416	0.270	0.401	1.779
Flood	0.223	0.146	0.230	0.423
Wildfire	0.205	0.156	0.219	0.627

TABLE 7. Classification performance on different datasets.

Method	BGS Dataset		Shipborne Dataset	
	RetNet	Ofli et al. [10]	RetNet	Li et al. [40]
Accuracy	**0.9160**	0.8700	0.8750	**0.9444**
Precision	**0.9067**	0.7370	**0.8852**	–
Recall	**0.9510**	0.6680	**0.9643**	0.8290
F1-Score	**0.9283**	0.7010	**0.9231**	–

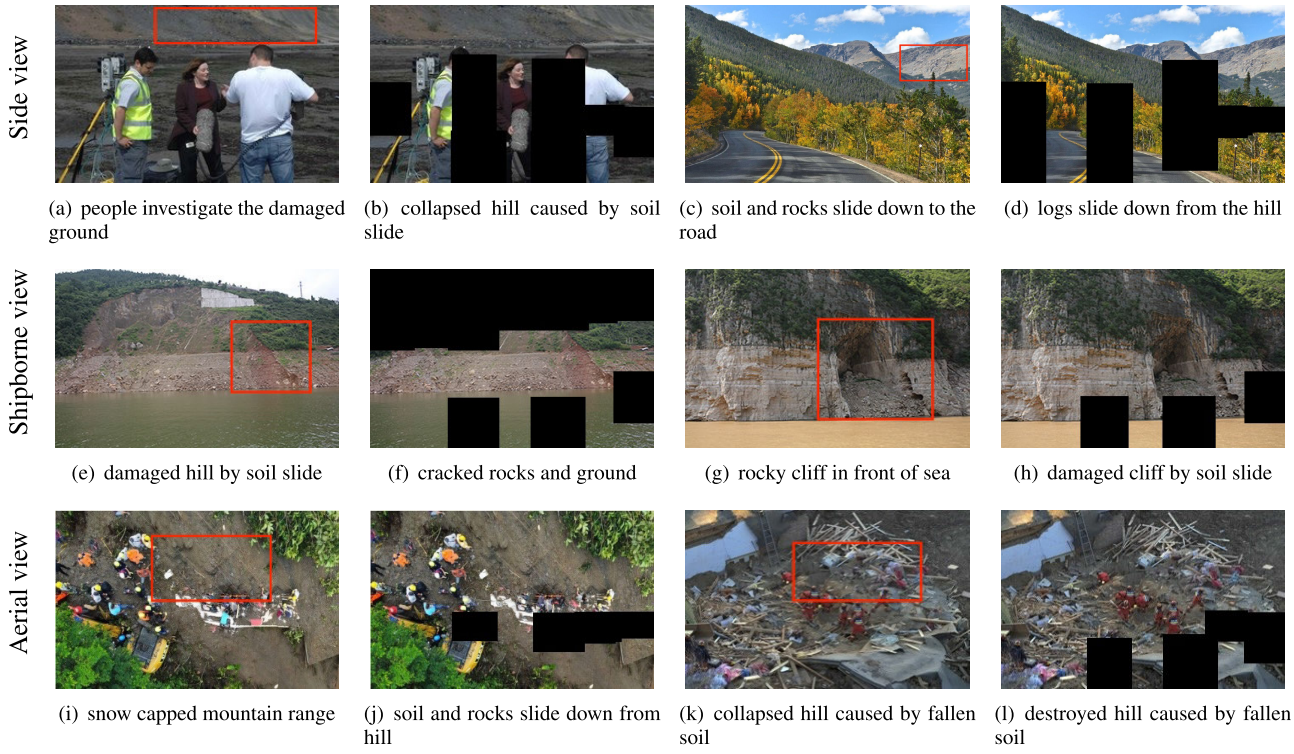
as shown in Table 5. Despite adjusting the number of anchor boxes in each grid and the patch grids, the modifications did not significantly impact the model's performance across all assessed metrics. The scores remained at 0.416, 0.339, 0.298, 0.258, 0.270, 0.401, and 1.799 for BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L, and CIDEr, respectively.

C. FLOOD AND WILDFIRE

We fine-tuned the model using images of floods and wildfires from the Diaster Image Dataset (DID). Descriptive captions were appended to each image for use in the training and testing stages. The results indicated that the model was more proficient at recognizing and captioning wildfire disasters compared to flood disasters, as shown in Figure 6. In particular, the BLEU-1 score for photographs of floods is 0.223, whereas that for images of flames was 0.205. Nevertheless, the ROUGE-L score of 0.230 for floods is marginally higher than the ROUGE-L score of 0.219 for wildfires, reflecting the matching of the longest common subsequence in captions of flood images. Moreover, the METEOR scores, 0.146 for floods and 0.156 for wildfires, demonstrate a reasonable level of semantic and syntactic agreement with reference-generated captions. The slightly higher score for wildfires suggests a superior model performance in describing wildfire imagery. Additionally, the METEOR and CIDEr scores—0.423 for floods and 0.627 for wildfires—reveal that the captions for wildfire imagery are more accurate and closely match the reference assessments.

D. CLASSIFICATION

The results of our comparative analysis of the classification performance of RetNet, VED, and ResNet50 [10] are listed in Table 7. As can be seen from the values of 0.9160,



(a) people investigate the damaged ground (b) collapsed hill caused by soil slide (c) soil and rocks slide down to the road (d) logs slide down from the hill

(e) damaged hill by soil slide (f) cracked rocks and ground (g) rocky cliff in front of sea (h) damaged cliff by soil slide

(i) snow capped mountain range (j) soil and rocks slide down from hill (k) collapsed hill caused by fallen soil (l) destroyed hill caused by fallen soil

FIGURE 6. Image captioning results vary significantly across different viewpoints, from side views as shown in Figure 6(a) to 6(d), shipborne views in Figure 6(e) to 6(h), and aerial views in Figure 6(i) to 6(l). Red boundary boxes highlight the disaster areas within the images.

0.9067, 0.9510, and 0.9283 for recall, accuracy, precision, and F1-score, respectively, our method performs better than the others. Furthermore, we used data from shipborne surveys to evaluate the classification performance from a different angle. Although RetNet's accuracy of 0.8750 was lower than that of the Fusion approach [40], RetNet outperformed the Fusion approach regarding F1-score, recall, and precision. Notably, our approach outperformed Fusion with a recall of 0.9643 against 0.8290 for Fusion.

VI. DISCUSSION

A. IMAGE CAPTIONING

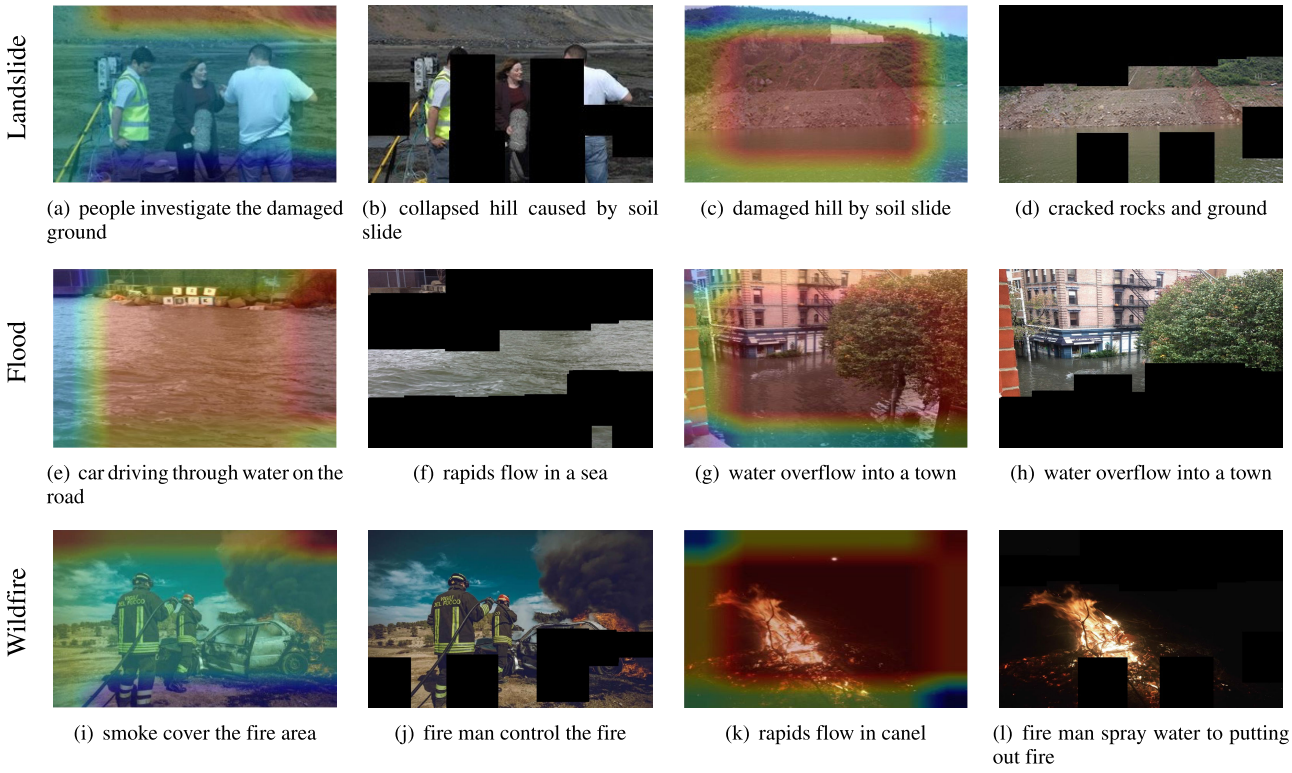
The results show that neither the patch size nor the number of anchor boxes significantly affect the generated captions, as shown in Table 3. The result was that the localization layer's area candidates were used to be candidates to be selected as the optimal ones for masking. The patch size and number of anchor boxes become insignificant considerations in image captioning since the masked regions frequently occupy similar positions.

Regarding the side-view images from Figure 6(a) to Figure 6(d), in contrast to the original captions, which might have concentrated on or misinterpreted other elements, our model prioritizes target objects more successfully. Furthermore, promising results of landslide detection from shipborne images, which are unseen datasets, are shown in Figure 6(e) to 6(h). In a similar vein, aerial images from Figs. 6(i) to 6(l) demonstrate the model's ability to identify landslides from an above-ground perspective. Even though the model is mainly trained on side-view images, our approach outperforms other approaches in landslide detection in aerial view images.

Additionally, we present a heat map illustrating the new attention derived from RetNet in Figure 7. For landslide disaster images, as shown in Figure 7(a), the heat map indicates that our model shifts its focus from people to the landslide region in the upper part of the image, thereby generating captions that describe the disaster situation in the targeted region. Conversely, in Figure 7(c), where the landslide occupies the center of the image without being obscured by other objects, the application of masking via RetNet enables the VED to concentrate exclusively on the specific landslide region depicted in the image.

In scenarios involving floods, the captions generated by our model accurately identified the specific areas depicted in the heat map images (Figure 7(e)). However, it is common for the original image (as shown in Figure 7(g)) and the corresponding masked image (as illustrated in Figure 7(h)) to produce identical captions. This occurrence is due to the extensive water regions in the images, which are sizable enough to be detected without the need for attention to shift.

For the wildfire disaster, Figure 7(i) and 7(j) show that our model not only allow VED to focus on wildfire but also on other surrounding objects and humans. As a result, the generated captions are correct even it is from different aspects from the original image. Moreover, in Figure 7(k), the



(a) people investigate the damaged ground
(b) collapsed hill caused by soil slide
(c) damaged hill by soil slide
(d) cracked rocks and ground
(e) car driving through water on the road
(f) rapids flow in a sea
(g) water overflow into a town
(h) water overflow into a town
(i) smoke cover the fire area
(j) fire man control the fire
(k) rapids flow in canel
(l) fire man spray water to putting out fire

FIGURE 7. Image captioning results vary significantly across different disasters, from landslide as shown in Figure 7(a) to 7(d), flooding in Figure 7(e) to 7(h), and wildrie in Figure 7(i) to 7(l). The heatmap red shown the focus of RetNet.

original image caption misinterpret the phenomena. On the other hand, our model could generate the correct perspective, but only partially correct captions as shown in Figure 7(l). In general, we realized that the VED captioning model would pay attention to the center of the image to generate the circumstance captions, while RetNet captions attention to the specific regions.

Finally, we highlight the cases of misdetection using our approach as shown in Figure 8. Specifically, RetNet mistakenly shifts the focus to the top regions of the side view of a fire disaster image (as shown in Figure 8(a)), rather than to the middle. This incorrect focus results in inaccurately generated captions. Similarly, in the side view image of a landslide (as shown in Figure. 8(b)), RetNet tends to focus on the upper part of the image rather than the correct target regions located at the center.

B. CLASSIFICATION

Target regions are the focus of our approach, as Table 7 shows that RetNet outperforms Ofli et al. [10] when it comes to side view image analysis, especially when using the BGS dataset. On the other hand, compared to the results of Li et al. [40], our model shows lower true positive and true negative rates when applied to shipborne images. In spite of this, RetNet achieves better recall, demonstrating its improved capacity to recognize landslides in uncertain scenarios.

Original: heavy fire is burning
RetNet: water overflow in residence

(a) Fire disaster failure case.

Original: stones slide donw to the ground
RetNet: buildings beside the canel with forest behind

(b) Landslide disaster failure case.

FIGURE 8. Heat map attention of caption from original image and RetNet image in failure case.

VII. CONCLUSION

Disaster-related areas often present a significant challenge for detection in aerial or shipborne imagery, primarily because of the small size of these regions within the images. In this study, we introduce a novel framework called the Retarget Network (RetNet), which is designed to enhance the ability of image-captioning-based machine learning models to pinpoint critical regions within an image.

Our proposed network uniquely adjusts detection priorities by integrating a localization layer with a retarget layer, using patch and anchor box techniques. We patched the image into square grid patches, and in each patch, anchor boxes were generated to serve as candidate regions in the localization layer. These candidates were then masked for shifted

attention in the image captioning model (VED) in the retarget layer. The ablation study found that the number of patches or anchor boxes in each patch did not affect the performance of the shifted attention in the image captioning model. Moreover, we found that the Inverse Cosine Similarity loss and L1 regularization outperformed the other combinations of loss functions. Inverse Cosine Similarity loss uses generated and reference captions to achieve the result, whereas L1 regularization penalizes parameter loss in RetNet to prevent model over-fitting and aids in feature selection from the image tokens.

The RetNet model was thoroughly tested across a range of disaster scenarios, including landslides, floods, and wildfires, from various perspectives. Our findings demonstrate that preprocessing images with RetNet before analysis with a Visual Explanation Detector (VED) significantly improves the accuracy of detecting landslides in side-view image captions to 91.60% and achieves an 87.50% accuracy rate for images captured from shipborne perspectives, which is an unseen dataset to evaluate the performance of RetNet in landslide disaster.

In this study, we found that complex images containing multiple main objects affected the RetNet's ability to shift attention in the image captioning model. Moreover, the caption tokens in this study were limited to 30 tokens due to the constraints of the labeled caption dataset, which impacted the precision of RetNet and the image captioning model (VED). These challenges, including handling complex scenes with multiple major objects and the limitation of text tokens, should be addressed in future work.

The limitation of RetNet with hazy or unclear images would make it quite challenging to retarget the major object if the surrounding objects have similar features. Moreover, the limitation of detection based on the image captioning model, VED, could also improve in the future.

REFERENCES

[1] O. Hungr, S. Leroueil, and L. Picarelli, "The varnes classification of landslide types, an update," *Landslides*, vol. 11, no. 2, pp. 167–194, Apr. 2014.

[2] N. S. Ibrahim, S. M. Sharun, M. K. Osman, S. B. Mohamed, and S. H. Y. S. Abdullah, "The application of UAV images in flood detection using image segmentation techniques," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 23, no. 2, p. 1219, Aug. 2021.

[3] D. Hernández, J. M. Cecilia, J.-C. Cano, and C. T. Calafate, "Flood detection using real-time image segmentation from unmanned aerial vehicles on edge-computing platform," *Remote Sens.*, vol. 14, no. 1, p. 223, Jan. 2022.

[4] H. Pan, D. Badawi, and A. E. Cetin, "Computationally efficient wildfire detection method using a deep convolutional network pruned via Fourier analysis," *Sensors*, vol. 20, no. 10, p. 2891, May 2020.

[5] M. Sardogan, A. Tuncer, and Y. Ozen, "Plant leaf disease detection and classification based on CNN with LVQ algorithm," in *Proc. 3rd Int. Conf. Comput. Sci. Eng. (UBMK)*, Sep. 2018, pp. 382–385.

[6] T. Trnovszky, P. Kamencay, R. Orjesek, M. Benco, and P. Sykora, "Animal recognition system based on convolutional neural network," *Adv. Electr. Electron. Eng.*, vol. 15, no. 3, pp. 517–525, Oct. 2017.

[7] S. R. Meena, L. P. Soares, C. H. Grohmann, C. Van Westen, K. Bhuyan, R. P. Singh, M. Floris, and F. Catani, "Landslide detection in the Himalayas using machine learning algorithms and U-Net," *Landslides*, vol. 19, no. 5, pp. 1209–1229, May 2022.

[8] L. P. Soares, H. C. Dias, and C. H. Grohmann, "Landslide segmentation with U-net: Evaluating different sampling methods and patch sizes," 2020, *arXiv:2007.06672*.

[9] P. Liu, Y. Wei, Q. Wang, Y. Chen, and J. Xie, "Research on post-earthquake landslide extraction algorithm based on improved U-Net model," *Remote Sens.*, vol. 12, no. 5, p. 894, Mar. 2020.

[10] F. Ofli, M. Imran, U. Qazi, J. Roch, C. Pennington, V. J. Banks, and R. Bossu, "Landslide detection in real-time social media image streams," 2021, *arXiv:2110.04080*.

[11] F. Ofli, U. Qazi, M. Imran, J. Roch, C. Pennington, V. Banks, and R. Bossu, "A real-time system for detecting landslide reports on social media using artificial intelligence," in *Web Engineering: 22nd International Conference, ICWE 2022, Bari, Italy, July 5–8, 2022, Proceedings*. Cham, Switzerland: Springer, Jul. 2022, pp. 49–65.

[12] H. Li, Y. He, Q. Xu, J. Deng, W. Li, and Y. Wei, "Detection and segmentation of loess landslides via satellite images: A two-phase framework," *Landslides*, vol. 19, no. 3, pp. 673–686, Mar. 2022.

[13] R. Can, S. Kocaman, and C. Gokceoglu, "A convolutional neural network architecture for auto-detection of landslide photographs to assess citizen science and volunteered geographic information data quality," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 7, p. 300, Jul. 2019.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[15] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[17] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4565–4574.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[20] N. Thanyawet, P. Ratsamee, Y. Uranishi, and H. Takemura, "Abnormal scene classification using image captioning technique: A landslide case study," in *Proc. IEEE 13th Int. Conf. Pattern Recognit. Syst. (ICPRS)*, Jul. 2023, pp. 1–7.

[21] M. I. Sameen and B. Pradhan, "Landslide detection using residual networks and the fusion of spectral and topographic information," *IEEE Access*, vol. 7, pp. 114363–114373, 2019.

[22] R. Castro, I. Pineda, W. Lim, and M. E. Morocho-Cayamcela, "Deep learning approaches based on transformer architectures for image captioning tasks," *IEEE Access*, vol. 10, pp. 33679–33694, 2022.

[23] S. Rothe, S. Narayan, and A. Severyn, "Leveraging pre-trained checkpoints for sequence generation tasks," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 264–280, Dec. 2020.

[24] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, "TrOCR: Transformer-based optical character recognition with pre-trained models," 2021, *arXiv:2109.10282*.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[26] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

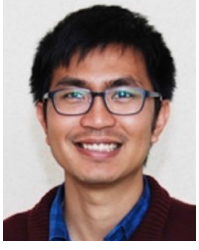
[28] S. Tantanee, K. Wandee, and S. Tovichakchaikul, "One page project management application on flood preparedness: Case study of Thailand," *Proc. Eng.*, vol. 212, pp. 363–370, Jan. 2018.

[29] R. S. Thomas and B. Hall, *Seawall Design*. London, U.K.: Butterworth, 2015.

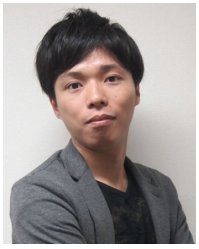
[30] I. Mirzaev, A. Yuvmitov, M. Turdiev, and J. Shomurodov, "Influence of the vertical earthquake component on the shear vibration of buildings on sliding foundations," in *Proc. E3S Web Conf.*, vol. 264, Les Ulis, France. EDP Sciences, 2021, p. 02022.

[31] M. Haseeb, A. B. Xinhailu, J. Z. Khan, I. Ahmad, and R. Malik, "Construction of earthquake resistant buildings and infrastructure implementing seismic design and building code in northern Pakistan 2005 earthquake affected area," *Int. J. Bus. Social Sci.*, vol. 2, no. 4, pp. 1–10, 2005.
[32] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 13041–13049.
[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[36] A. Tanatipuknon, P. Aimmanee, Y. Watanabe, K. T. Murata, A. Wakai, G. Sato, H. V. Hung, K. Tungpimolrut, S. Keerativittayanun, and J. Karnjana, "Study on combining two faster R-CNN models for landslide detection with a classification decision tree to improve the detection performance," *J. Disaster Res.*, vol. 16, no. 4, pp. 588–595, Jun. 2021.
[37] (2023). *The National Archive of Geological Photographs*. [Online]. Available: http://geoscenic.bgs.ac.uk/asset-bank/action/viewHome
[38] (2020). *Landscape Pictures*. [Online]. Available: https://www.kaggle.com/datasets/arnaud58/landscape-pictures
[39] (2019). *Disaster Image Dataset*. [Online]. Available: https://www.kaggle.com/datasets/mikolajbabula/disaster-images-dataset-cnn-model
[40] Y. Li, P. Wang, Q. Feng, X. Ji, D. Jin, and J. Gong, "Landslide detection based on shipborne images and deep learning models: A case study in the Three Gorges Reservoir Area in China," *Landslides*, vol. 20, no. 3, pp. 547–558, Mar. 2023.
[41] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[42] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.
[43] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.
[44] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.

NARONGTHAT THANYAWET (Member, IEEE) received the B.E. and M.E. degrees in engineering from Chulalongkorn University, Thailand, in 2020. He is currently pursuing the Ph.D. degree with the Graduate School of Information Science and Technology, Osaka University, Osaka, Japan. His research interests include computer vision, image processing, and natural language processing, including their application in environments and disasters.



PHOTCHARA RATSAMEE (Member, IEEE) received the M.E. and Ph.D. degrees from the Graduate School of Engineering Science, Osaka University, in 2012 and 2015, respectively. He was an Assistant Professor with the Cybermedia Center, Osaka University, from 2015 to 2022. He is currently an Associate Professor (Lecturer) with the Faculty of Robotics and Design, Osaka Institute of Technology. His research interests include robot vision, rescue robots, human–robot interaction, and haptics in mixed reality. He is a member of the Robotics Society of Japan (RSJ) and the IEEE Robotics and Automation Society (RAS).



YUKI URANISHI (Member, IEEE) received the Ph.D. degree from the Graduate School of Information Science, Nara Institute of Science and Technology, Japan, in 2008. He is currently a Professor with the Cybermedia Center, Osaka University, Japan. Before joining Osaka University, he was a Research Fellow with Japan Society for the Promotion of Science and an Assistant Professor with Nara Institute of Science and Technology, from 2009 to 2012; Osaka University, from 2012 to 2014; and Kyoto University Hospital, from 2014 to 2016. He was an Associate Professor with the Cybermedia Center, Osaka University, from 2016 to 2024. His research interests include computer vision, augmented reality, virtual reality, and human–computer interaction.

MASATO KOBAYASHI (Member, IEEE) received the Bachelor of Maritime Sciences, Master of Maritime Sciences, and Doctor of Engineering degrees from Kobe University, Japan, in 2017, 2019, and 2022, respectively. From 2019 to 2021, he was an Engineer and a Researcher with the Technology Development Division, Seiko Epson Corporation, Japan. From 2021 to 2022, he was a Research Internship with OMRON SINIC X Corporation, Japan. Since 2022, he has been an Academic Researcher with Kobe University. He was with the Yutaka Matsuo Laboratory, The University of Tokyo, researching and developing robotics and AI. Since 2023, he has been with Osaka University, Japan, where he is currently an Assistant Professor. His current research interests include robotics, motion control, haptic, XR, and AI.

HARUO TAKEMURA (Life Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from Osaka University, Osaka, Japan, in 1982, 1984, and 1987, respectively. He was a Researcher with the Advanced Telecommunication Research Institute, International (ATR-I), Kyoto, Japan, from 1987 to 1994; and an Associate Professor with Nara Institute of Science and Technology, Nara, Japan. He has been a Professor with the Infomedia Education Division, Cybermedia Center, Osaka University, since 2001. He is in charge of campus wide deployment of learning management system (LMS) and other IT systems for education. His research interests include interactive computer graphics, human–computer interaction, mixed reality, and their applications in education, including learning analytics.