Citation: A. Sarkar, T. Chowdhury, R. Murphy, A. Gangopadhyay and M. Rahnemoonfar, "SAM-VQA: Supervised Attention-Based Visual Question Answering Model for Post-Disaster Damage Assessment on Remote Sensing Imagery," in IEEE Transactions on Geoscience and Remote Sensing, doi: 10.1109/TGRS.2023.3276293.

DOI: https://doi.org/10.1109/TGRS.2023.3276293

SAM-VQA: Supervised Attention-Based Visual Question Answering Model for Post-Disaster Damage Assessment on Remote Sensing Imagery

Argho Sarkar, *Member, IEEE*, Tashnim Chowdhury, *Member, IEEE*, Robin Roberson Murphy, *Fellow, IEEE*, Aryya Gangopadhyay, *Member, IEEE*, and Maryam Rahnemoonfar, *Member, IEEE*

***Abstract*— Each natural disaster leaves a trail of destruction and damage that must be effectively managed to reduce its negative impact on human life. Any delay in making proper decisions at the post-disaster managerial level can increase human suffering and waste resources. Proper managerial decisions after any natural disaster rely on an appropriate assessment of damages using data-driven approaches, which are needed to be efficient, fast, and interactive. The goal of this study is to incorporate a deep interactive data-driven framework for proper damage assessment to speed up the response and recovery phases after a natural disaster. Hence, this article focuses on introducing and implementing the visual question answering (VQA) framework for post-disaster damage assessment based on drone imagery, namely supervised attention-based VQA (SAM-VQA). In VQA, query-based answers from images regarding the situation in disaster-affected areas can provide valuable information for decision-making. Unlike other computer vision tasks, VQA is more interactive and allows one to get instant and effective scene information by asking questions in natural language from images. In this work, we present a VQA dataset and propose a novel SAM-VQA framework for post-disaster damage assessment on remote sensing images. Our model outperforms state-of-the-art attention-based VQA techniques, including stacked attention networks (SANs) and multimodal factorized bilinear (MFB) with Co-Attention. Furthermore, our proposed model can derive appropriate visual attention based on questions to predict answers, making our approach trustworthy.**

***Index Terms*— Attention, post-disaster management, remote sensing, search and rescue, visual question answering (VQA).**

I. Introduction

DISASTER management can be defined as an accountable organization and management for dealing with all humanitarian aspects, particularly post-disaster response and recovery, to mitigate the impact of a disaster. In the response and recovery stage after any catastrophic event, disaster management requires a fast and interactive data-driven approach to thoroughly comprehend the damaged situation. A rapid and in-depth understanding of the damage in the aftermath of disasters is essential for supporting the decision-making system. The decisions regarding the distribution of relief and food to the highly victimized areas, the operation of the search and rescue missions, the reconstruction of the damaged roads and buildings, etc., are dependent on the proper assessment of the damage. Any delay in the recovery phase can drive human lives toward death and dissipate an abundance of money. Haas et al. [3] established a logarithmic heuristic which suggests that reducing the time spent on each phase of a disaster response reduces the time spent on the next phase by a factor of 10. In this article, we present a supervised attention-based visual question answering (SAM-VQA) framework to provide high-level scene information for proper damage assessment to speed up the response and recovery phases after any natural disaster.

Visual question answering (VQA) is a complicated multimodal research problem in which the aim is to answer an image-specified question. In a VQA framework, we generally ask questions about images in natural language. Thus, the VQA framework needs to model the question and visual content to get the most appropriate answers from images. Substantial research efforts have been made on the VQA task in the computer vision and natural language processing communities [1], [4], [5], [6], [7] using deep learning-based multimodal methods. The key benefit of the VQA method is that it can promptly deliver high-level scene information from images through interaction, which is limited in other computer vision tasks. Image segmentation [8], [9], [10], [11] segments an image into several object categories, objection detection [12], [13] algorithms detect objects from an image. However, these tasks do not consider both providing high-level scene information and interacting with users. On the other hand, in any VQA task, a model needs to detect objects (object detection), classify their attributes (classification), and figure out the interactive relationship among different entities within images to provide answers. This high-level scene understanding has the potential to advance the post-disaster managerial decision support system, especially in the rescue mission. In this interactive framework, a non-domain expert can obtain information regarding damages from images by

Manuscript received 4 November 2022; revised 25 March 2023; accepted 29 April 2023. Date of publication 15 May 2023; date of current version 2 June 2023. This work was supported in part by the U.S. Army under Grant W911NF2120076 and in part by the Microsoft and Amazon. *(Corresponding author: Maryam Rahnemoonfar.)*

Argho Sarkar, Tashnim Chowdhury, and Aryya Gangopadhyay are with the Center for Real-time Distributed Sensing and Autonomy (CARDS), University of Maryland Baltimore County, Baltimore, MD 21250 USA.

Robin Roberson Murphy is with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843 USA.

Maryam Rahnemoonfar is with the Department of Computer Science and Engineering and the Department of Civil and Environmental Engineering, Lehigh University, Bethlehem, PA 18015 USA (e-mail: maryam@lehigh.edu).

Digital Object Identifier 10.1109/TGRS.2023.3276293

asking questions in natural language. Fig. 2 presents how rescuers can utilize the VQA task in the post-disaster damage assessment. This direct interaction also makes this data-driven approach faster compared to other data-driven approaches. In this study, VQA for post-disaster damage assessment has been considered on the extension of the FloodNet-VQA dataset proposed in [14]. This new version of FloodNet-VQA V2.0 has more types of image-question pairs related to damages after a hurricane. "How many buildings are flooded?," "Is the road flooded?," "Do the rescuers need to provide help urgently?," and "How is the building density in this area?" are some examples. Answers to those questions certainly provide a deep understanding of the condition of the affected areas to the rescuers, which assists them in estimating the damage and providing direction to take action. This motivates us to include the task of VQA for post-disaster damage assessment.

Existing VQA algorithms [1], [4], [7] are mostly trained on ground-based images. However, in this research, we consider an aerial imagery-based (e.g., drone) VQA framework. The concept of developing drone-based VQA stems from the characteristics of a drone which is its ability to reach remote areas for data collection during or after any natural disaster. However, developing a drone imagery-based VQA algorithm for post-disaster damage assessment is extremely difficult for many reasons. First, the representation of drone images refers to a top-down (vertical) view, which is different compared to the human-centric (horizontal) representation captured by traditional digital cameras. Top-down pictorial representations make it very difficult to distinguish between several objects, as the objects of interest become relatively small. Second, in the case of damage assessment, the degree of scene complexity gets much higher due to noises coming from many sources, such as structural debris. Therefore, special care needs to be taken in the modeling part to successfully provide correct answers from the drone-based VQA system.

Attention-based VQA models [1], [2], [7], [15], [16] showed remarkable performance on many ground imagery-based VQA datasets. Attention in VQA algorithms is defined as assigning weights within different image regions according to the importance of getting clues for predicting the answer to a given question. Relevant image portions should get higher weights compared to irrelevant portions to answer a question. Although those attention-based VQA frameworks can obtain relevant visual attention weight from many ground imagery-based VQA datasets, they fail to obtain relevant visual attention from remote sensing images. The main reason for not obtaining relevant visual attention weights on remote sensing images is the way those models are learning visual attention weights. Most of the attention-based VQA models are trained in a supervised manner (i.e., minimizing the cross-entropy loss between the ground truth and predicted answers). However, visual attention weights within those models are learned without any additional supervision and solely based on *ground-truth answers*. By ground-truth answer, we mean the corresponding true text answer to a given question about an image. The estimated visual attention weight distributions for remote sensing images learned solely by minimizing loss between the *ground-truth answers* and predicted answers in a classification manner could not properly highlight the relevant image regions. An additional learning component as a means of supervision is needed so that the estimated visual attention weight can focus on relevant image portions to answer a question. Thus, to supervise the visual attention weight, we need the visual ground-truth which will highlight the relevant image portions necessary for answering a given question. To address this, we propose a SAM-VQA framework to obtain relevant visual attention weights on remote sensing images in the context of post-disaster damage assessment. In contrast to existing approaches, our proposed approach allows *visual attention* to be supervised by the *visual mask* equivalent to visual ground-truth along with the supervision of the model by ground-truth answers. The *visual mask* is generated from the image based on the corresponding question. For example, if the question is about the road, then the *visual mask* is generated by masking all other parts of the corresponding image except the "road." Thus, a visual mask provides attention weight distribution over an image by highlighting the relevant visual portions based on the question. Our approach will allow learning from both the visual mask and true answer distribution jointly. The visual mask allows the model to learn relevant visual attention weights and the true answer distribution enables the model to predict rational answers. We name this process of providing supervision to images as *visual supervision*. Fig. 1 compares the quality of the derived visual attention map of the SAM-VQA approach with the approach that does not consider additional visual supervision for remote sensing images. It is clearly shown that when the estimated visual attention weights are learned through visual supervision, the algorithm learns better where to put attention on the image

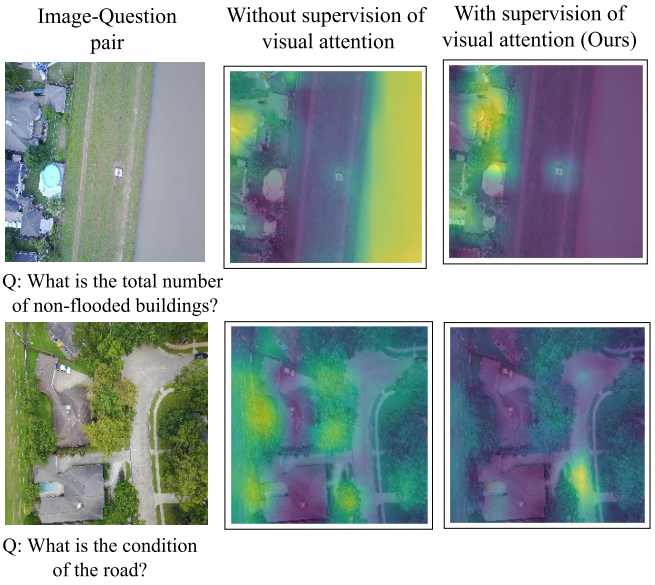

Fig. 1. Comparison of derived visual attentions for given questions from two VQA models, one of which is trained without visual supervision and the other with visual supervision. The yellowish tone in the image denotes higher attention weight. Attention learned with visual supervision (the last column) emphasizes the relevant image portions (buildings and roads in this case) to address the questions from the top and bottom images, respectively. On the other hand, the attention learned without visual supervision (the middle column) fails to pay proper attention to both images.

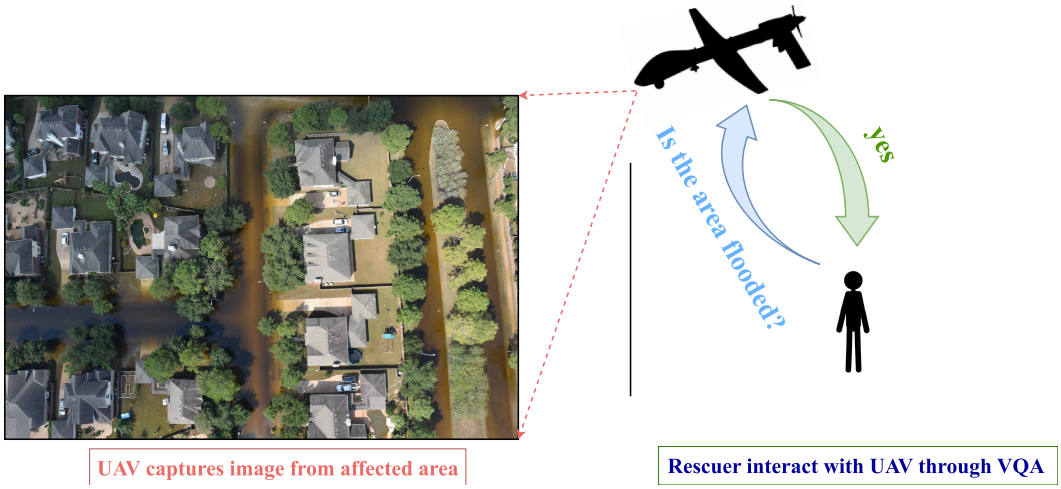

Fig. 2. Rescuer can acquire effective information about the affected area by asking questions when a drone coupled with a VQA system captures images from the hurricane-stricken area from a high altitude.

content to predict the answer compared to the process where attention is learned without visual supervision. For example, to answer the questions from the top and bottom images in Fig. 1, the model must emphasize the building and road portions of the images, respectively. When we trained the model without visual supervision, it could not pay attention to the corresponding relevant image portions. Fig. 9 demonstrates our proposed architecture.

The main contributions of this research work consist of the following.

1) Develop a VQA dataset, *FloodNet-VQA V2.0*, which is the extension of *FloodNet-VQA* [14] for post-disaster damage assessment purposes.
2) Propose a novel SAM-VQA on top of the developed dataset.
3) We experimentally showed that our approach is more accurate in terms of providing correct answers and trustworthy in respect of providing relevant visual attention compared to the state-of-the-art attention-based VQA methods.

The organization of this article is as follows. In Section II, existing works on natural disaster assessment and VQA are discussed. Section III provides the details on the dataset. Our newly proposed method and the result are described in Sections V and VI, respectively. Finally, the future work and the conclusion have been addressed in Sections VII and VIII, respectively.

II. Related Works

We will discuss the notable works done for natural disaster damage assessment based on aerial and satellite imagery in the first subsection. The latter subsection discusses research on VQA for remote sensing images.

A. Natural Disaster Damage Assessment

Most of the research on natural damage assessment is limited to structural damage detection (e.g., detecting damaged buildings), classification (e.g., classifying the level of damage associated with structures), and semantic segmentation. In recent time, many aerial [10], [14], [17], [18] and satellite imagery [8], [9], [13], [19], [20], [21] have been proposed for the aforementioned computer vision tasks. Feng et al. [22] propose a model to estimate the risk of causality based on damages to buildings after a disaster. In [12], collapsed buildings were detected from the aerial images after earthquakes. In [23], structural damage assessment is conducted based on multiperspective, overlapping, very high-resolution oblique images obtained with unmanned aerial vehicles (UAVs). In [18], AIST Building Change Detection (ABCD) aerial dataset has been proposed that includes post-tsunami images to investigate if the buildings have been washed away. Chen et al. [13] propose a dataset including aerial and satellite imagery in order to detect building damage after a hurricane. A segmentation model is proposed to identify the structural-level changes and estimate the effects of natural disasters in [8]. Aerial Image Database for Emergency Response (AIDER) is proposed in [17], which aims to classify UAV imagery. A dataset collected from both Sentinel-1 and Sentinel-2 satellites is introduced in [9]. This dataset offers semantic segmentation of flooded buildings. In order to determine building damages, xBD is proposed in [19], which includes both pre- and post-event satellite images. UAV-based datasets and deep learning approaches have been proposed in [14], [24], and [25] for post-disaster damage assessment purposes after hurricanes. Many approaches such as image segmentation [24], [26], [27] and instance segmentation [10] are proposed, where images are collected using UAVs. Tilon et al. [28] conduct experiments to detect damaged buildings using the xBD dataset. In [29], a semi-supervised technique is proposed to detect the damaged building based on satellite images.

B. Vision-Language Model for Remote Sensing

Recently, vision- and language-based multimodal approaches (e.g., image captioning, visual question generation, and VQA) are gaining attention in the remote sensing

community. Many datasets are provided for image captioning tasks, including RSCID [30], UCM-Captions [31], and Sydney-Captions [31]. A summarization-driven deep remote sensing image captioning algorithm is proposed in [32]. The authors in this work integrated the summary of the caption with ground-truth captions to overcome information redundancy as captions are repetitive or semantically similar to each other. Scene attention, defined as utilizing both the semantic information from long short-term memory (LSTM) and the global visual information from features to generate an attention map, is proposed in [33] for image captioning tasks through an encoder–decoder-based architecture. Visual question generation (VQG), one of the vision-language-based multimodal tasks, is proposed in [34], where the motivation is to generate meaningful questions from remotely sensed images.

Besides the above tasks, remarkable progress has been made in VQA for remote sensing. Many works have been proposed to tackle this challenge. In [35], two VQA datasets for remote sensing, in general, have been proposed. Furthermore, a large-scale remote sensing VQA dataset has been proposed in [36] which includes 15 million image-question-answer (QA) triplets from the BigEarthNet dataset. Different fusion strategies between image and text features for fine-grained multimodal feature extraction for the VQA task have been studied in [37]. Attention-based VQA frameworks are also explored in much of the research for remote sensing. The mutual attention network [38] considers both the convolutional feature map and the semantic visual feature vector from the image model, as well as the question vector from the question model, to achieve mutual attention. This joint representation is further fed into a fully connected (FC) layer for answer prediction. A cross-modal attention-based VQA method has been proposed in [39] for remote sensing. In this cross-modal technique, image and question representations are fed to a cross-modal transformer network that uses cross-attention between the image and text modalities to generate the answer. Besides remote sensing in general, several application-based VQA methods have recently been proposed. In [14] and [40], study of VQA algorithms for post-disaster damage assessment has been carried out. Change detection in the form of VQA has been proposed in [41].

In this research work, we proposed a SAM-VQA framework for post-disaster damage assessment on the extended FloodNet-VQA V2.0 dataset. Our research aims to provide high-level scene information through interaction following any disaster to advance the decision support system.

III. FloodNet-VQA V2.0 Dataset for Post-Disaster Assessment

Like other deep learning models, the success of VQA also relies on a large volume of image data. There are various sources after a disaster from which the image data can be obtained. Human participation is involved in most of those traditional data-collection processes. Due to a variety of adverse circumstances during natural disasters, such as damaged highways, flooded areas, and so on, human involvement in the data collection process is very risky in terms of safety. Drones are an efficient way of collecting images from impacted areas without the need for human intervention. Our VQA framework, in this study, is built for drone imagery. In this section, we will discuss the process of image collection, data annotation, and the development of the VQA dataset. At the end of this section, we discuss the difference between FloodNet-VQA and its extension, FloodNet-VQA V2.0.

A. Image Collection

The data collection process took place after the *Hurricane Harvey*. *Hurricane Harvey* was a Category 4 hurricane that hit Texas and Louisiana in August 2017, causing catastrophic flooding and killing over 100 people. We take advantage of the UAV platform to capture images and videos from the affected areas. DJI Mavic Pro quadcopters have been used for the data collection process. The data were collected after conducting several flights covering areas mostly in Ford Bend County, Texas, and other directly impacted areas between August 30 and September 4, 2017. Fig. 4 represents the risk level in many counties in Houston, Texas. The dataset is unique for two reasons. First, images are very high in resolution, and second, it is the only known database for small UAV (sUAV) disaster imaging. In all the images, the post-flood damages from the affected areas are depicted. Other collections of imagery, utilizing unmanned and manned aerial assets captured during disasters, such as the National Guard Predators or Civil Air Patrol are larger and fixed-wing assets that operate above the 400 feet above ground level (AGL). Our images, taken from a height of 200 feet, have a very high spatial resolution of about 1.5 cm, making them special among natural disaster datasets. We consider a fixed size of 4000×3000 for all the images. All available aerial and satellite images [8], [9], [10], [13], [18], [19], [20] vary in size and have a lower resolution than ours. Our high-resolution imagery provides detailed information, which leads to a good understanding of the situation.

B. Image Annotation Task

QA pairs are generated from the pixel-wise annotations (i.e., semantic segmentation) of images [14]. We annotate each image pixel-wise to identify multiple objects and their attributes. Fig. 8 shows some examples of pixel-wise annotation. Building-flooded, building-non-flooded, road-flooded, road-non-flooded, water, tree, car, pool, and grass are the nine classes that have been assigned pixel-wise labels in each of the images. When at least one side of a building is in contact with flood water, it is classified as a flooded building. To differentiate between natural and flood water, the "water" class has been created, which represents any natural water body like a river or lake. In addition, each image is labeled as "flooded" or "non-flooded" as a whole. If flood water covers more than 30% of an image, it is classified as a flooded image; otherwise, it is classified as a non-flooded image. After annotation, we created a dictionary for each image that stores each object type and corresponding annotation details (e.g., polygon). The multiple presences of the same object in a particular annotation dictionary indicate the frequency of that

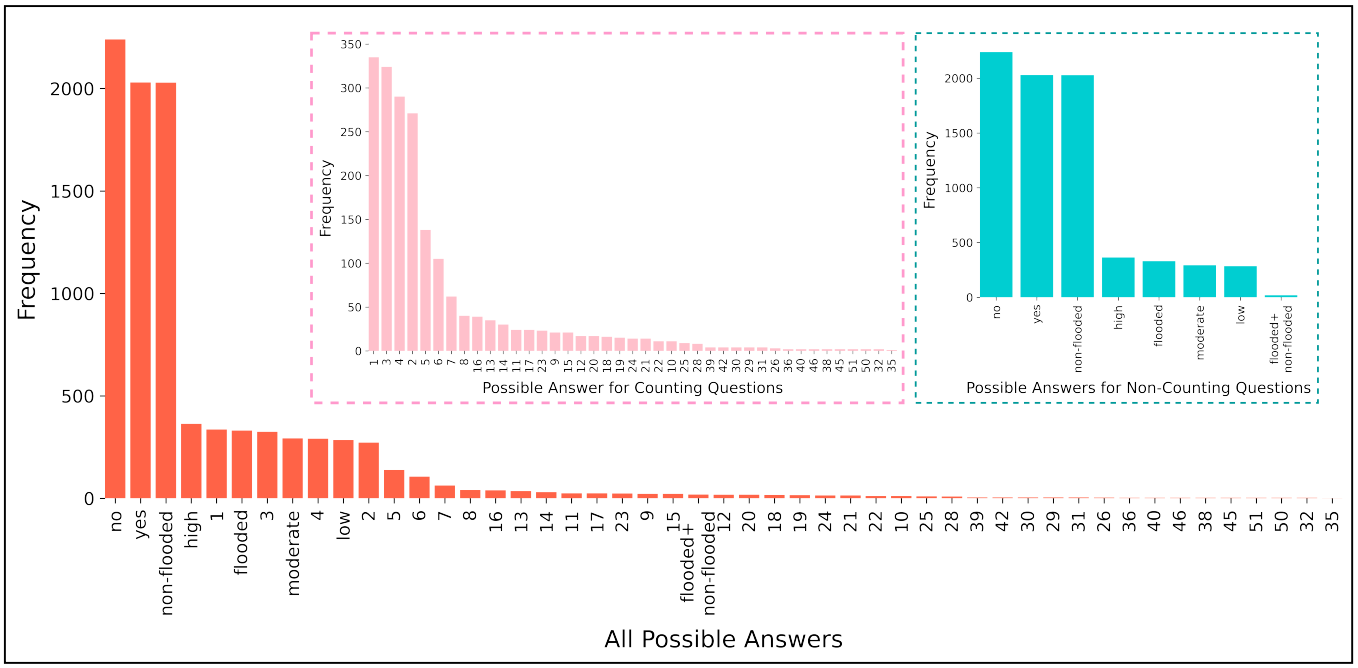

Fig. 3. Distribution of answers from FloodNet-VQA V2.0. The outer figure represents the distribution of all answers from the dataset. The nested left and right figures show the distribution of answers for counting and non-counting questions, respectively.

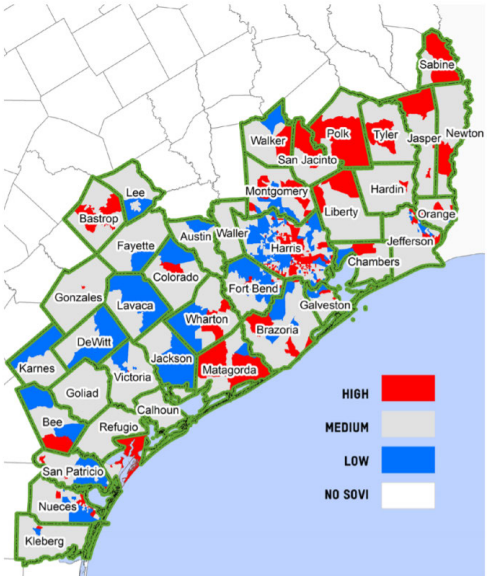

Fig. 4. Risk level among counties in Houston due to Hurricane Harvey [42].

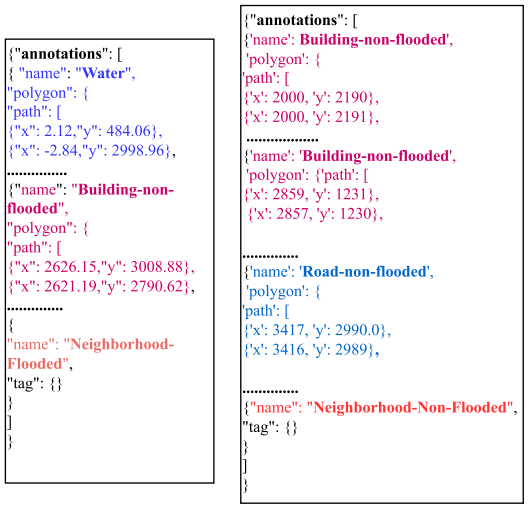

Fig. 5. Example of image dictionaries for two different images. Each dictionary contains annotation information regarding the objects and corresponding attributes.

object in that particular image. Fig. 5 provides an example of the annotation dictionary. This dictionary is used to generate ground-truth answers for all the questions in our dataset.

C. Data Quality Assurance

Throughout the annotation process, we have maintained the quality of pixel-level annotations. A two-stage quality assurance system has been followed. Each annotation must be approved by reviewers, who decide the quality of the annotations by ensuring that annotators follow the rules established by experts in the field. If the annotation is not up to par, reviewers send it back to the same annotator with a list of issues and suggestions for improvement. The annotations that were rejected are then re-annotated and forwarded to the reviewers for approval. This cycle is repeated until all of the images have been properly annotated and the standards have been accurately followed.

D. Dataset Preparation

1) Question Category Selection: The selection of question categories is very important so that the information from the questions can allow the rescuers to have a better understanding of damages to the affected areas. For making decisions at the post-disaster managerial level, a lot of information needs to be analyzed so that the recovery process can be made faster and

Fig. 6. Distribution of question categories where each slice in the pie chart represents the ratio between the number of questions in a particular category and the total number of questions.

more efficient. The recovery process includes the distribution of manpower engaged in rescue missions, the distribution of relief supplies, and so on. These factors are taken into account while generating the questions. Each question in the dataset provides different spectra of information through answers to enhance the decision support system. The distribution of different question categories in our dataset is presented in Fig. 6.

1) One of the objectives of understanding the post-disaster scene is to identify image-level characteristics (i.e., the flooded and non-flooded images). We design questions that will provide image-level information through answers. "What is the overall condition of the entire image?" is an example of this question category. To generate ground-truth answers for this question category, we follow the rules described in Section III-B. We define this question category as *Entire Image Condition Recognition*.
2) Rescuers need to identify the condition of roads to save lives and operate their search and rescue mission. This recognition allows the rescuers to see whether the impacted area is reachable by road before they start their move to that specific area. "What is the condition of the road?," "Is the road flooded?" will definitely serve this purpose. This question category is denoted as *Road Condition Recognition*.
3) Identifying building-level damage is also very important in the rescue mission. To address this, we include *Building Condition Recognition* question category. "Is there any flooded buildings?" is an example of this kind.
4) Counting the structural entities provides an intuition of the level of risk and damage in an area due to the disaster. For instance, if numerous buildings in a specific location are damaged by a disaster, the extent of the damage will be higher in that area compared to places where fewer buildings are damaged. We separate these counting-related questions into two categories, namely simple counting and complex counting. In the *Simple Counting* problem, we ask about an object's frequency of presence (mainly buildings) in an image regardless of the attribute (e.g., "How many buildings are in the image?"). *Complex Counting* is specifically intended to count the number of a particular building attribute (e.g., "How many **flooded or non-flooded** buildings are in the image?"). We are interested in counting only the flooded or non-flooded buildings under this category of question. In comparison to simple counting, a higher level of scene understanding can be obtained by complex counting.
5) Identifying the level of density of the structures will help in the rescue mission by distributing the limited manpower. Highly dense areas need to allocate more rescuers than less dense areas in the recovery process. Hence, the proposed dataset includes *Density Estimation* question category. "What is the building density of the area?" is an example of this category.
6) In the recovery process, rescuers should take immediate action in highly affected areas as the level of risk associated with human life at that location is high. For that, rescuers need to identify the level of risk in different affected areas. To serve this purpose, we include *Risk Assessment* question category. "Do the rescuers need to provide help urgently in this area?," "Does this area need immediate help?" are some examples from this question category.

2) Answer Generation: Fig. 5 represents the example of an annotation dictionary used in this study. Dictionaries contain the information from which answers to the questions are assigned. We have followed some rules or thresholds while generating answers.

1) To identify whether an image is flooded, we consider the value from *Neighborhood* key in the corresponding image dictionary. *Neighborhood* key in the annotation dictionary contains the image-level information (flooded or non-flooded).
2) Fig. 5 contains information regarding the road condition, namely road-flooded and road-non-flooded. This information is used to assign answers for the questions related to *Road Condition Recognition*.
3) Answers to counting questions are generated by counting the presence of a building with or without an attribute from the corresponding image dictionary. For example, in Fig. 5 (right side), building-non-flooded appears twice. Thus, the answer to the question "How many buildings are non-flooded?" will be 2.
4) The level of density can be identified from the number of buildings in an image. In this study, we consider an area highly dense if the number of buildings is greater than 5. If the number of buildings is between 3 and 5, we consider it moderate, and if the frequency of buildings is lower than 3, we consider it a less dense area. As the density depends on the image resolution, the setting of these threshold levels for density estimation is applicable only to our dataset. Our images, taken from a height of 200 feet, have a very high spatial resolution of about 1.5 cm.
5) Answers to the questions in the *Risk Assessment* question category depend on the number of flooded buildings

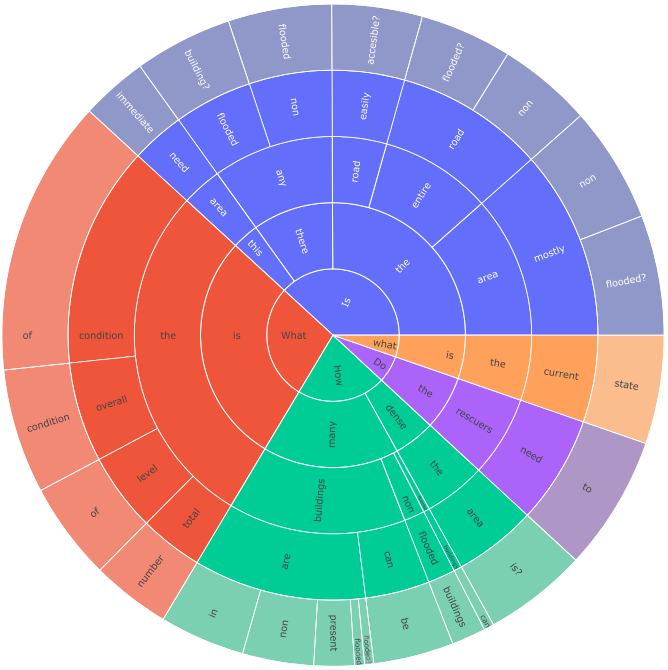

Fig. 7. Distribution of the questions by their first four words. This plot highlights the variety of questions in FloodNet-VQA V2.0. The length of an arc is in proportion to the number of questions involving that word.

present in an image. This is because the greater the number of flooded buildings, the more risky the area is. The threshold point (the frequency of flooded buildings in a given image) for taking immediate action is 3.

Fig. 3 refers to the distribution of possible answers of our dataset.

E. FloodNet-VQA V2.0 Versus FloodNet-VQA

Extension of *FloodNet-VQA*, *FloodNet-VQA V2.0*, includes 2348 images and 10 480 QA pairs. On the other hand, *FloodNet-VQA* includes 7355 QA pairs. The distribution of the number of questions for each category of *FloodNet-VQA V2.0* is presented in Fig. 6. In addition, this new extension includes three more types of question categories, including *Density Estimation*, *Risk Assessment*, and *Building condition Recognition*. Due to the necessity of comprehending the damaged scenario completely and making the rescue mission effective, these types of questions are incorporated into the dataset. *Road Condition Recognition* and *Entire Image Condition Recognition* include the highest number of questions. The number of words in the longest question is 11 for our dataset. Fig. 7 presents the distribution of questions' varieties based on the starting word. Questions starting with the word "what" have the highest variation compared to other starting words.

Table I compares the data statistics between *FloodNet-VQA* and *FloodNet-VQA V2.0*.

TABLE I
COMPARISON BETWEEN FLOODNET-VQA AND FLOODNET-VQA V2.0 FOR DIFFERENT ATTRIBUTES

	FloodNet-VQA	**FloodNet-VQA V2.0**
Number of images	2188	2348
Number of questions	7355	10,480
Number of question type	4	7
Number of unique questions	15	43
Maximum word count in questions	11	11
Average word count in questions	7.89	7.93
Number of unique answers	41	49

IV. GENERATION OF VISUAL MASK (GVM)

Generally, "visual attention" is defined as giving importance to relevant image regions (i.e., pixels) for a prediction. In VQA, understanding the specific visual content (e.g., objects, relations among different objects, or attributes of objects) within an image is important for providing answers. The task of attention in VQA is to prioritize image portions that are highly relevant to answering a question. The attention in VQA is question-dependent, meaning different visual attentions are required to answer different questions from the same image. Failing to focus on the proper image regions (incorrect attention weight) leads to wrong answers. This is because matrix multiplication between the image feature matrix and the estimated attention vector, shown in Fig. 9, is fed into the VQA classifier to predict the answer. As a result, a misleading attention weight vector can manipulate matrix multiplication and be responsible for the wrong prediction for a given image-question pair. Thus, we include the visual mask highlighting the relevant visual clues in the training process so that the model understands where to direct more visual attention for a given question in order to predict the answer.

To generate visual masks, we mask the irrelevant portions of the images based on the questions. From the questions, we first identify which objects or regions within images are important for providing correct answers. Then we leverage the annotation from semantic segmentation to mask images. From semantically segmented images, we mask the irrelevant

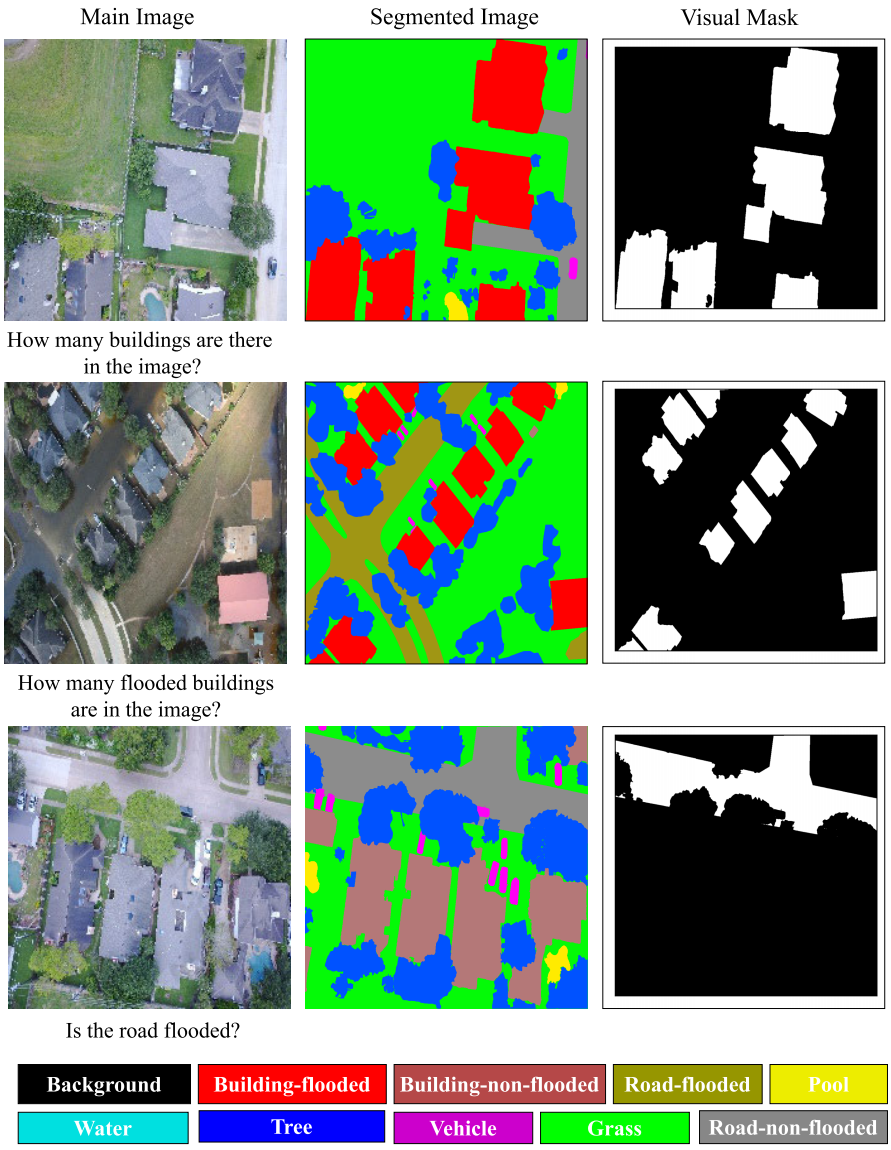

Fig. 8. Overview of the dataset. Each image is associated with the corresponding semantically segmented image and visual mask. These visual masks provide supervision to the visual attention obtaining process, enabling our model to learn where it should focus.

objects or portions of images by replacing the pixel value with [0, 0, 0], considering the RGB channel, and highlight the relevant portions by replacing the pixel values with [1, 1, 1]. The process of obtaining the visual mask for each question category is described below.

1) In *Building Condition Recognition* question category, the answer will be either "yes or no." For example, if the question is "Is there any flooded buildings?" and the corresponding ground-truth answer is "no," then the model needs to pay more attention to the image regions related to non-flooded buildings, and if the answer is "yes," the model needs to focus on parts of flooded buildings. These flooded buildings and non-flooded buildings are categorized by the level of water described in Section III. Thus, we generate the visual mask by only highlighting the non-flooded buildings for the first case and the flooded buildings for the second case.
2) In *Complex Counting*, the model needs to count attribute-specific buildings. To count the flooded buildings in an image, the model needs to pay attention to the regions where flooded buildings are located. On the other hand, to count non-flooded building models, the model needs to pay higher attention to non-flooded buildings in an image. As a result, if the question is "How many flooded buildings are there in the image?" we only highlight flooded buildings, and if the question is "How many non-flooded buildings are there in the image?" we only highlight non-flooded buildings. The second example from the top in Fig. 8 shows the visual mask of this question category.
3) In *Simple Counting*, we highlight all the buildings regardless of attributes (e.g., flooded or non-flooded) and mask the rest. The first example from the top in Fig. 8 presents the visual mask of this question category.

4) In the *Density Estimation* question category, the model needs to count the buildings, regardless of attributes, to give an answer. As a result, the focus should be placed on all the buildings depicted in the image. To generate visual masks for this question category, we only highlight all the buildings and mask the rest of the portions of the images.
5) In *Entire Image Condition Recognition* question category, two types of answers are involved. One type of answers describes the situation, and another type provides an answer in binary ("yes/no") form. For instance, answers (flooded or non-flooded) related to the question "What is the overall condition of the image?" provide situational information. On the other hand, the answer to the question "Is the area mostly flooded?" will be between "yes and no." To generate visual masks for this question category, we relied on the answers. If the question is "What is the overall condition of the image?" and the related answer is "flooded," the model should pay attention to flooded regions such as flooded buildings or flooded roads. Thus, the visual mask in this case is generated by highlighting the flooded buildings and flooded road-related image regions. On the other hand, if the question is "Is the area mostly flooded?" and the corresponding answer is "no," then we generate the visual mask by masking out the whole image except the non-flooded buildings and non-flooded road-related regions, and vice versa.
6) In *Risk Assessment* question category, attention should be given to the flooded buildings in the images to provide answers. This is because the more flooded buildings there are in a location, the more people living in that area are in danger. Thus, we only highlight the flooded buildings from images while generating the visual masks.
7) Like *Entire Image Condition Recognition*, *Road Condition Recognition* includes two types of answers. To generate visual masks for this question category, we also relied on the answers. If the answers are "flooded" and "yes" for the questions "What is the condition of the road?" and "Is the road easily accessible?" respectively, we highlight the flooded roads and non-flooded roads. Otherwise, we highlight the non-flooded and flooded roads. The last image from the top in Fig. 8 represents an example of the visual mask for this question category.

V. SAM-VQA Model

Due to the challenges involved in the drone imagery-based VQA approach, described in Section I, visual attention weight estimated without additional visual supervision fails to give importance to the most relevant image portions. Therefore, we propose to provide visual supervision through the visual mask for better estimating visual attention weight on drone images along with ground-truth answers. Fig. 9 represents our proposed VQA model, in which we provide the visual mask by masking the irrelevant image portions (i.e., regions in the image that are not necessary to look at for predicting an answer) for a given question. In this way, visual attention weights can be learned by minimizing the distance between the visual mask and estimated visual attention distribution (namely, attention loss) along with categorical cross-entropy loss between the ground-truth and predicted answers. It is worth mentioning that answers in our VQA approach are predicted in a classification manner. By learning this auxiliary loss (attention loss), our proposed VQA approach is able to learn to focus on the relevant image portions for a given question. As a result, we enhance the performance of our proposed VQA algorithm for post-disaster damage assessment and obtain relevant visual attention maps. Though we train our proposed VQA model with two losses, we only consider the classification part to predict answers in the test phase. In this section, we will discuss the components of our proposed SAM-VQA model.

A. Problem Formulation

Let our dataset $\mathcal{D}$ has n number of sample data: $\{(\mathcal{I}_1, \mathcal{Q}_1, \mathcal{GVM}_1), (\mathcal{I}_2, \mathcal{Q}_2, \mathcal{GVM}_2), \ldots, (\mathcal{I}_n, \mathcal{Q}_n, \mathcal{GVM}_n\}$. Here $\mathcal{I}$, $\mathcal{Q}$, and $\mathcal{GVM}$ refer to the main image of interest, the corresponding question, and visual mask, respectively. The objective of VQA is to predict the answer $\hat{a}$ from a set of possible answers $\mathcal{A}$ to the given question $\mathcal{Q}$ from the image $\mathcal{I}$

$$\hat{a} = \arg\max_{a \in \mathcal{A}} f_\theta(a|\mathcal{I}, \mathcal{Q}).$$

Here, f is the learnable model with $\theta \in \Theta$ trainable parameters.

B. Proposed VQA Framework

Our VQA framework depends on four important steps:

1) Visual Feature Extraction
2) Question Feature Extraction
3) Fusion of Visual and Question Features
4) Visual Attention Derivation

1) Step-1: Visual Feature Extraction: We first obtain the image feature matrix $f_\mathcal{I}$, described in Fig. 9, from the last pooling layer of the CNN (Resnet-152) architecture. At first, we resize the RGB images to be $224 \times 224 \times 3$ and then extract image feature matrices of size $14 \times 14 \times 1024$, where 14×14 (=196) is the number of grids in an image feature matrix and 1024 is the dimension of each grid feature vector. In another way, each grid in the feature matrix represents the 16×16 image region of an input image

$$f_\mathcal{I} = \text{ResNet}(\mathcal{I}) \in \mathbb{R}^{m \times n \times d}.$$

Here, m, n, and d represent the height, width, and number of channels of the image feature matrix, respectively. In this case, value for m, n and d are 14, 14, and 1024, respectively.

2) Step-2: Question Feature Extraction: For question-level feature representation, two-layer LSTM has been taken into account. To extract the semantic features of the question, we obtained the feature vector $f_\mathcal{Q}$ from the last cell of the last layer of the LSTM. We considered the dimension of the question feature vector to be 1024

$$f_\mathcal{Q} = \text{LSTM}(\mathcal{Q}) \in \mathbb{R}^d.$$

Here, d represents 1024 dimensional question feature vector.

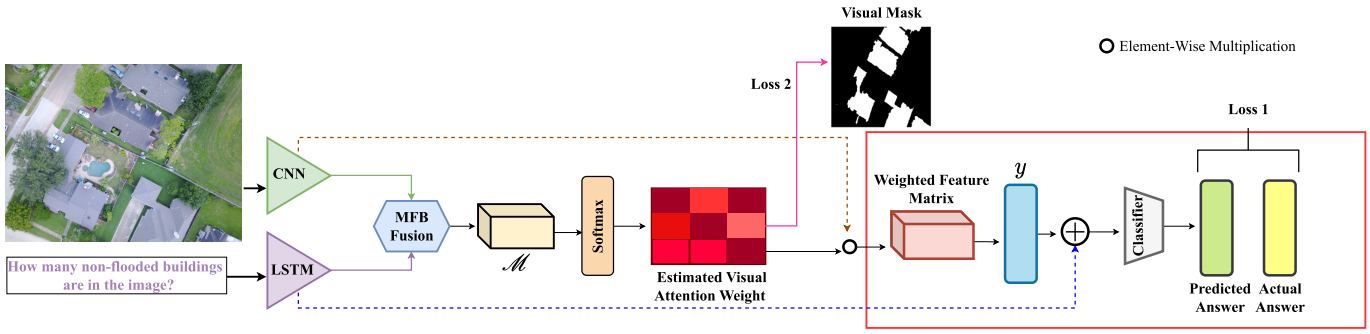

Fig. 9. Overview of our proposed SAM-VQA model. In this framework, we use Resnet-152 and a two-layer LSTM to obtain the image feature matrix and question feature, respectively. We then consider MFB pooling to obtain a fine-grained multimodal representation. A softmax function is applied to that joint representation to estimate attention weights from the images for given questions. Finally, we calculate two loss functions: one minimizes the distance between the visual mask and the estimated visual attention weight, and the other minimizes the loss between the ground-truth answer and the predicted answer from the VQA classifier.

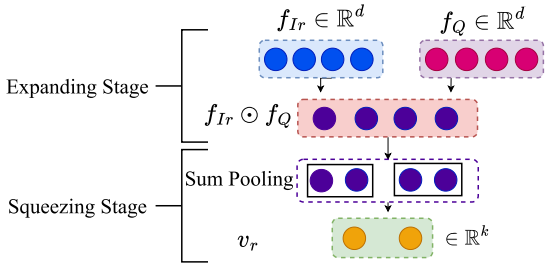

Fig. 10. Workflow of MFB pooling technique.

3) Step-3: Fusion of Visual and Question Features: In Fig. 9, we represent that each image grid feature vector $f_{\mathcal{I}r} \in \mathbb{R}^d$, where $r \in \{1, 2, \ldots, mn\}$ represents the grid index, is fused with the corresponding text feature vector $f_{\mathcal{Q}} \in \mathbb{R}^d$ using the concept of multimodal factorized bilinear (MFB) pooling. MFB pooling is shown in Fig. 10. Given the feature vectors from two modalities, image grid feature vector $f_{\mathcal{I}r} \in \mathbb{R}^d$ and question feature vector $f_{\mathcal{Q}} \in \mathbb{R}^d$, the fusion strategy is divided into two stages. In the expanding process, the image grid and question feature vectors are multiplied element-wise, followed by a dropout layer as follows:

$$m_r = f_{\mathcal{I}r}^T \odot f_{\mathcal{Q}} \in \mathbb{R}^d.$$

In the squeezing step, sum pooling is considered, followed by power and ℓ_2 normalization layers. $\odot$ refers to element-wise multiplication operation

$$v_r = \ell_2(\text{power}(\text{SumPool}(m_r))) \in \mathbb{R}^k$$

where the notation $\text{SumPool}(x, w)$ denotes the sum pooling operation over x using a 1-D, non-overlapping window of size w. In this case, $w = 2$ is set. k is the new output dimension after sum pooling which is 512 in our study. The power normalization ($m_r \leftarrow sign(m_r)|m_r|^{0.5}$) and $\ell_2(m_r \leftarrow m_r/||m_r||)$ layers are added after MFB output. $\mathcal{M} = (\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{mn}}) \in \mathbb{R}^{m \times n \times k}$ is the final output matrix considering MFB operation for all r.

4) Step-4: Visual Attention Derivation: The visual attention derivation process is shown in Fig. 9. The *Softmax* function is applied to the $\mathcal{M}$, from Step-3, to estimate the visual attention weight $p_r, r \in \{1, 2, 3, \ldots, mn\}$, for each mn image grids from the image feature matrix $f_{\mathcal{I}}$. Let us define $p \in \mathbb{R}^{mn}$ as the estimated attention weight vector. In another way, this estimated visual attention weight vector can be defined by $\widehat{\mathcal{GVM}}$ which is supervised by $\mathcal{GVM}$ in our proposed model. Each value in the estimated attention weight vector refers to the contribution of that corresponding image grid to the prediction for a corresponding question. After the attention layer, two branches are considered. One branch minimizes the distance between the visual mask and estimated visual attention weight, namely *Loss 2* and the second branch considers the classification task of predicting answers. In the second branch, each image grid feature vector $f_{\mathcal{I}r}$ is multiplied by the corresponding estimated attention weight p_r to generate a weighted feature matrix. The weighted feature vector (or attention vector) y is then the sum of the weighted feature matrix across each grid. This weighted feature vector is fed into the classifier for the prediction by minimizing the *Loss 1*. We consider categorical cross-entropy loss as *Loss 1* and Kullback–Leibler divergence (KL loss) as *Loss 2*

$$y = \sum_{r=1}^{mn} f_{\mathcal{I}r} p_r. \tag{1}$$

Here, y is the attention vector calculated by taking the sum of the weighted feature matrix over each grid r, where $r \in \{1, 2, 3, \ldots, mn\}$ represents the grid index.

C. Classification

Weighted feature vector y is summed up with $f_{\mathcal{Q}}$ and fed into the FC layer and finally, the output from the FC layer is fed into the Softmax layer for answer prediction $\hat{a}$

$$\hat{a} = \text{Softmax}(\text{FC}(y + f_{\mathcal{Q}})). \tag{2}$$

D. Loss Function

The joint loss function for the proposed model is

$$L_{\text{total}} = -\sum_{i=1}^{|\mathcal{A}|} a_i \log P(a_i|\mathcal{I}, \mathcal{Q}) + \beta KL(\mathcal{GVM}||\widehat{\mathcal{GVM}}). \tag{3}$$

Here, $|\mathcal{A}|$ is the length of possible answers. The first term is categorical cross-entropy loss and the second term is KL divergence loss. On the other side, $\mathcal{GVM}$ and $\widehat{\mathcal{GVM}}$ are the visual mask and estimated visual attention weight vector, respectively. β is the scaling parameter.

VI. Results and Discussion

A. Model Comparison

Five baseline models are compared to our proposed model. There are two types of models: attention-free (*VIS + LSTM*, *CNN + LSTM*) and attention-based [*Stacked Attention Network (SAN)*, *MFB Pooling with Co-Attention* (MFB + CoAtt)].

1) *Question-Only:* In this experiment, we only provide questions to predict answers. This model addresses the effect of language bias in VQA.
2) *VIS + LSTM:* This is a simple, attention-free VQA approach [43] where the image is fed into a convolutional net (CNN) and the LSTM is considered to predict the answer. In this algorithm, the LSTM is initialized with the output from the CNN.
3) *CNN + LSTM:* In this attention-free approach [44], image features are extracted by CNN, and question features are extracted from the last cell of one-layer LSTM. Finally, these features are fused by element-wise multiplication and fed into the MLP layer to predict the answer.
4) *Stacked Attention Networks (SAN):* SAN [1] is a state-of-the-art attention-based model for VQA. In this approach, multistep attention is considered to predict the answer. In our experiment, we adopted the SAN model with two-step attention.
5) *MFB Pooling With Co-Attention (MFB + CoAtt):* This is another state-of-the-art model [2] that considers co-attention mechanisms to predict the answer. In co-attention, both word-level attention from a question and image-level attention are computed. However, prior to guiding the image feature to derive attention over image regions, question attention is derived. In the fusion stage, image and question features are fused with MFB pooling.

B. Implementation Details

We consider the batch size to be 32, and the *adam* optimizer is used during the training phase. The learning rate for this study is set to 0.001, and the learning rate of each parameter group is decayed by a factor of 0.1 every five epochs. Image feature is extracted from the *Resnet-152* model, and a *two-layer LSTM* model is chosen for question feature extraction. Models are trained from scratch, meaning that we have not considered pre-trained weights for image and text (e.g., Glove embedding). We have considered 6000 image-question pairs for the model training purposes, 1800 QA samples for the validation, and finally 2680 samples for testing purposes.

C. Ablation Study

1) Effect of Different Visual Encoders: The visual encoder plays an important role in VQA to extract meaningful image feature, which gets refined further after being fused with the corresponding question feature. The more relevant features are extracted, the more prediction of a model is rational. To investigate the effect of different visual encoders on model accuracy, we evaluated three CNN algorithms, namely VGG-16, Resnet-50, and Resnet-152. To compare the effects of different visual encoders, we have considered three combinations: 1) visual encoders with one-layer LSTM; 2) visual encoders with two-layer LSTM; and 3) visual encoders with three-layer LSTM.

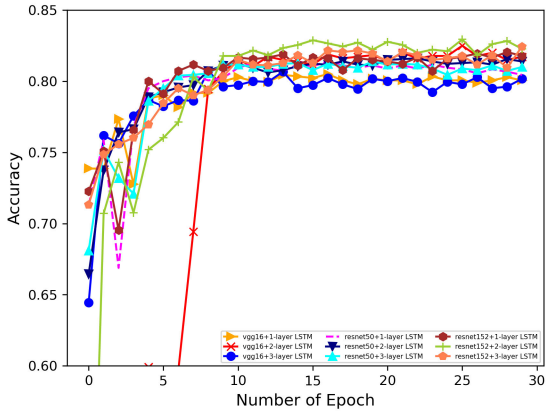

Fig. 11. Performance of different visual encoders on model accuracy for the validation data.

The comparison is shown in Fig. 11. From that figure, we can see that there is less variation in prediction accuracy between different visual encoders. However, Resnet-152 performs better compared to VGG-16 and Resnet-50.

2) Effect of Different Language Encoders: To study the effect of different language encoders, we have examined the impact of one-layer, two-layer, and three-layer LSTM on different visual encoders. From Fig. 12, we can see that the performance of the two-layer LSTM is slightly better than that of the one-layer and three-layer LSTM for all visual encoders. For VGG-16, this difference is much higher compared to Resnet-50 and Resnet-152.

3) Effect of Visual Supervision: To prove the acceptability of our proposed visual supervision technique, we have considered several combinations of different visual and language encoders: 1) VGG-16 with one-layer, two-layer, and three-layer LSTMs; 2) Resnet-50 with one-layer, two-layer, and three-layer LSTMs; and 3) Resnet-152 with one-layer, two-layer, and three-layer LSTMs. These combinations are then considered for two experimental settings: 1) train each of these models without visual supervision and 2) train each of these models with visual supervision. A comparison of the accuracy on the validation data between models without supervision and models with supervision is provided in Fig. 13. We found that, in all combinations, the model with supervision outperformed the model without supervision. This proves our hypothesis that providing visual supervision in the training stage improves accuracy. The highest accuracy from 30 epochs for all these models on the validation dataset is represented in Table II. Table III compares the accuracy between the models with and without visual supervision on the test dataset. We can

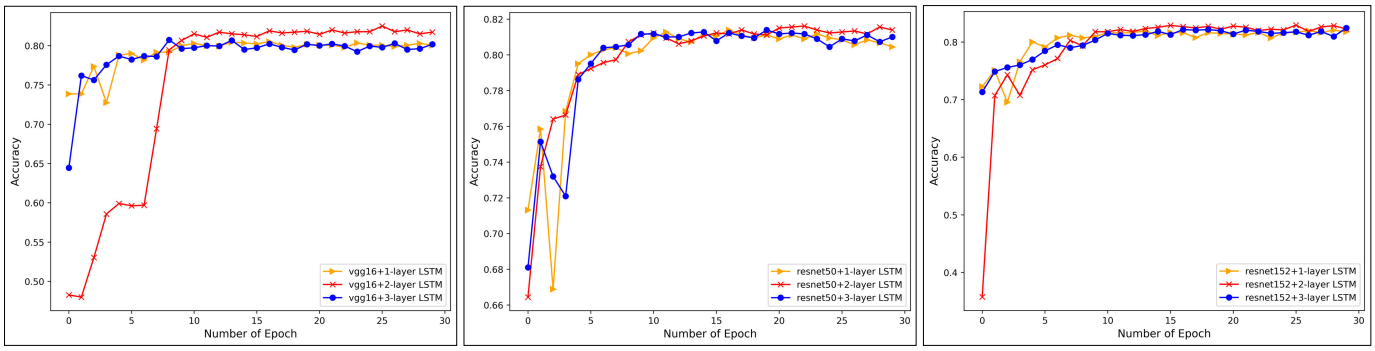

Fig. 12. Performance of different language encoders on model accuracy for the validation dataset. 1) The left figure depicts the effect of language encoders when VGG-16 is used as an image encoder. 2) The middle figure shows the effect of language encoders while considering Resnet-50 as an image encoder.3) The right figure shows the effect of language encoders while considering Resnet-152 as an image encoder.

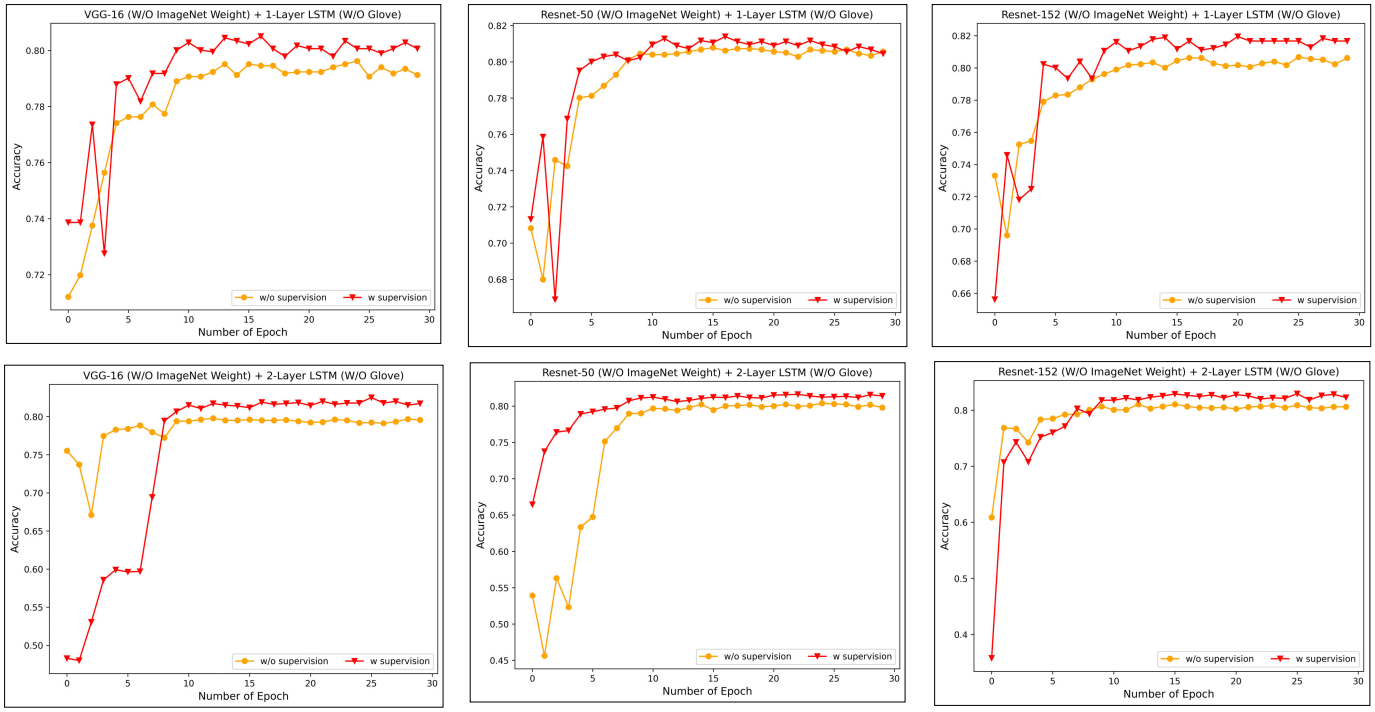

Fig. 13. Effect of the proposed visual supervision technique of our SAM-VQA model on the validation dataset. Several combinations of visual and language encoders are taken into consideration when comparing the effectiveness of visual supervision with that of not having visual supervision. In all combinations, the model with visual supervision outperforms the model without visual supervision.

TABLE II

EFFECT OF VISUAL SUPERVISION ON THE VALIDATION DATA

		Without Visual Supervision	With Visual Supervision (ours)
VGG-16	1-Layer LSTM	0.796	**0.805**
	2-Layer LSTM	0.797	**0.800**
	3-Layer LSTM	0.795	**0.807**
Resnet-50	1-Layer LSTM	0.808	**0.814**
	2-Layer LSTM	0.804	**0.816**
	3-Layer LSTM	0.811	**0.820**
Resnet-152	1-Layer LSTM	0.807	**0.821**
	2-Layer LSTM	0.811	**0.829**
	3-Layer LSTM	0.803	**0.821**

TABLE III

EFFECT OF VISUAL SUPERVISION ON THE TEST DATA

		Without Visual Supervision	With Visual Supervision (ours)
VGG-16	1-Layer LSTM	0.77	**0.78**
	2-Layer LSTM	0.78	**0.79**
	3-Layer LSTM	0.78	**0.79**
Resnet-50	1-Layer LSTM	0.77	**0.78**
	2-Layer LSTM	0.77	**0.78**
	3-Layer LSTM	0.79	**0.80**
Resnet-152	1-Layer LSTM	0.78	**0.79**
	2-Layer LSTM	0.78	**0.81**
	3-Layer LSTM	0.79	**0.80**

identify that, in both cases, the model with visual supervision outperforms the model without visual supervision.

D. Accuracy Assessment

Table IV shows the comparison of the model accuracy between the baseline methods and our proposed method. We consider top-1 accuracy in our study. From the table, we can see that our proposed SAM-VQA outperforms all the baseline methods. The overall accuracy of our model is 0.81. This accuracy is 8% higher than the *Question-only* model. *Question-only* model shows that the model can predict the answer with higher accuracy from the question itself

TABLE IV

ACCURACY COMPARISON BETWEEN SAM-VQA AND OTHER BASELINE MODELS ON THE TEST DATASET FOR DIFFERENT QUESTION CATEGORIES

VQA-Model	Simple Counting	Complex Counting	Road Condition	Building Condition	Density Estimation	Risk Assessment	Entire Image	Overall
Question-only	0.14	0.14	0.82	0.81	0.25	0.79	0.88	0.63
VIS + LSTM [43]	0.16	0.14	0.97	0.92	0.55	0.94	0.97	0.75
CNN + LSTM [44]	0.22	0.22	0.97	0.93	0.63	0.94	0.97	0.77
MFB+CoAtt [2]	0.27	0.27	0.98	0.91	0.68	0.97	0.98	0.79
SAN [1]	0.30	0.28	0.98	0.93	0.68	0.97	0.98	0.80
SAM-VQA (ours)	**0.35**	**0.32**	**0.98**	**0.94**	**0.74**	**0.97**	**0.98**	**0.81**

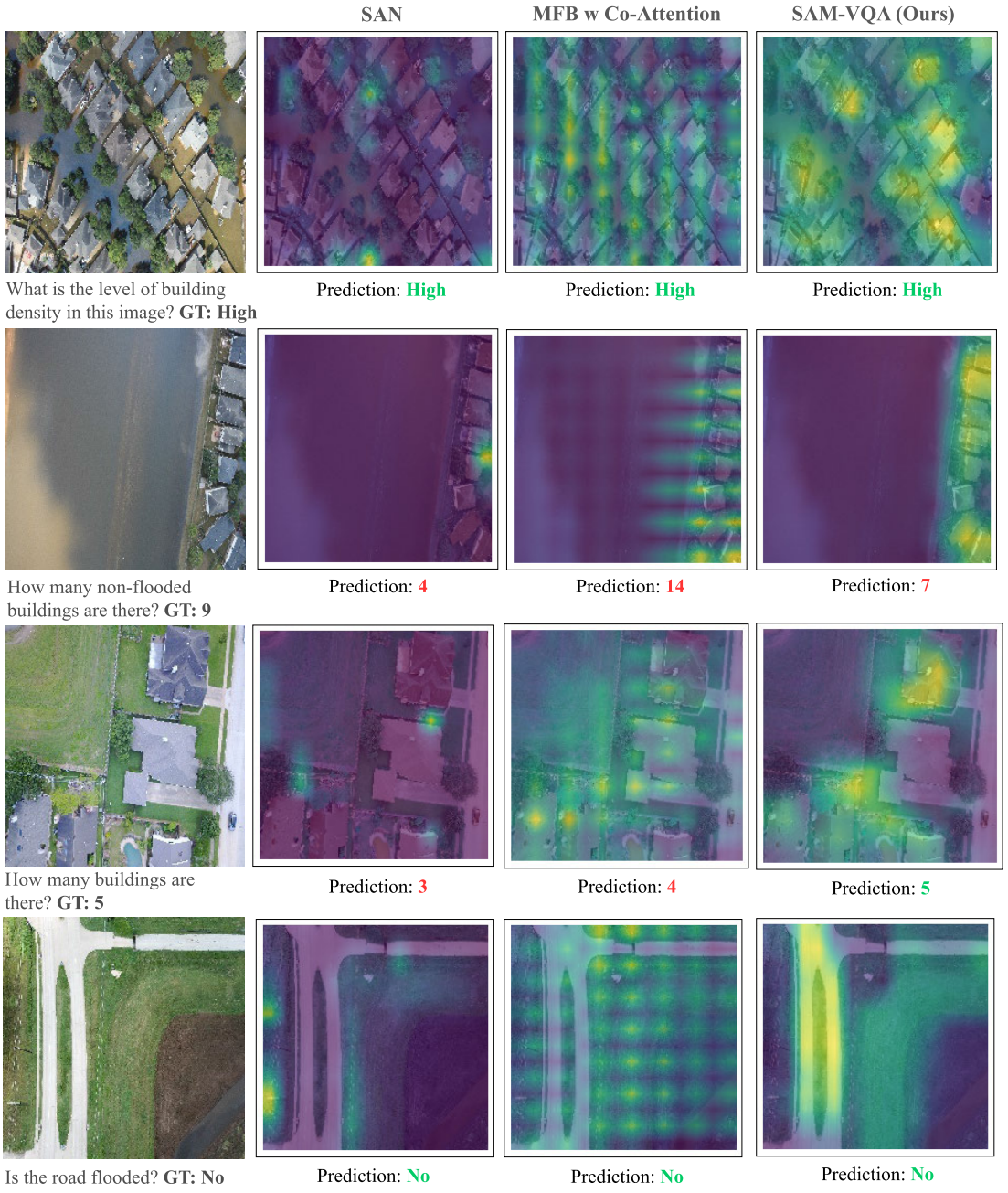

Fig. 14. Comparison of the visual attention map of the proposed SAM-VQA model with other baseline models for given image-question pairs. Correct prediction is indicated by green, and wrong prediction is indicated by the red word color. The higher the attention, the darker the transition from a green color to a yellow color in the image.

for *Entire Image Condition Recognition* and *Risk Assessment* question categories. This is due to the imbalanced distribution of answers from these two categories. Fig. 3 depicts this imbalanced distribution. In this experiment, we compare our model with two attention-free baseline models, namely *VIS + LSTM* [43] and *CNN + LSTM* [44]. Our model

TABLE V
COMPARISON OF MSE BETWEEN THE ESTIMATED VISUAL ATTENTION WEIGHT AND VISUAL MASK ON THE TEST DATASET

	MSE
Without Visual Supervision	0.0924
With Visual Supervision (ours)	**0.0912**

outperforms those two models by a larger margin. The capability of our proposed model can be identified when we compare it with state-of-the-art attention-based models. The overall accuracy of our proposed model is 2% and 1% higher than the MFB + CoAtt [2] and SAN [1] models, respectively.

From the results mentioned in Table IV, we understand that providing correct answers regarding the counting and *Density Estimation* question categories are very challenging. However, our *SAM-VQA* model is more accurate in providing the correct answers for these categories, while other baseline models struggle a lot. Our proposed model exceeds the accuracy by a margin of 5% and 4% in *Simple* and *Complex Counting* question categories, respectively compared to the most competitive *SAN* model. Based on the results from the *Simple* and the *Complex counting* question categories, we can further interpret that by adding visual supervision in the training phase, the proposed model is able to differentiate between flooded and non-flooded buildings more accurately. On the other hand, the *SAM-VQA* model outperforms in the *Density Estimation* question category by 6% compare to the attention-based baseline models. For the rest of the question categories, our SAM-VQA model also outperforms other baseline models.

E. Quality of Derived Visual Attention Map

The main purpose of our proposed SAM-VQA approach is to obtain relevant visual attentions from given questions. Proper attention makes the VQA model trustworthy. Our proposed VQA pipeline is capable of drawing relevant visual attention. Fig. 14 visually proves that attention in our model is much more relevant compared with the other two attention-based VQA approaches, namely SAN and MFB with Co-Attention. To answer the question "What is the level of building density in this image?" from the top image in Fig. 14, a trustworthy model needs to focus on the building regions in the image. Our proposed *SAM-VQA* model highlights that image portions properly, whereas the SAN model fails to provide that proper visual attention. However, predictions for that image-question pair from all three models are correct. For *Complex Counting* question, the second image from the top in Fig. 14, our proposed model perfectly provides the relevant visual attention, buildings in this case, whereas the attention from MFB with Co-Attention is sparse. Although the predictions are incorrect for all three models, our proposed SAM-VQA model predicts the building number with the least amount of error compared to the other two approaches. In the last example from Fig. 14, we further see that our model can identify the road from the image and provide the correct prediction to the question related with road, whereas the other two models predict the answer correctly but fail to provide relevant visual attention. To have a quantitative analysis of the visual attention map, we consider the mean squared error (MSE) between the visual mask and the estimated visual attention vector from our proposed model. In Table V, we see that our proposed model achieves a lower MSE compared to the baseline model (model without visual supervision). Based on the quantitative and qualitative results presented above, we can conclude that supervising visual attention during the training phase improves the accuracy and reliability of our proposed model.

F. Language Bias

A major issue with VQA models is language bias, which occurs when the prediction of the model relies mainly on the question rather than the image. For each question in the test set, we randomly selected an image to examine the language bias in our model. In the test set, we find an overall accuracy of 77.10%. This slight decline in accuracy from 81% in the test data suggests that there is less language bias and the model mostly extracts information from images based on questions.

VII. FUTURE WORK

In this study, we propose a novel VQA framework on remote sensing images to assess post-disaster damage. For the objective of visual supervision, we manually generate the visual mask. Manual annotation is time-consuming and expensive, so our future work focuses on developing a supervision system that eliminates the need for manual annotation.

VIII. CONCLUSION

In this study, we present the concept of VQA for post-disaster damage assessment purposes. We highlight the importance of the VQA task in damage assessment after any natural disaster. From this study, we provide a VQA dataset in the context of post-disaster damage assessment and develop a novel SAM-VQA algorithm. Our experiment demonstrated that our proposed model is more accurate in providing answers to the questions than the state-of-the-art attention-based baseline models. Finally, we showed that providing visual supervision substantially increases the model's reliability by obtaining proper visual attention that is relevant to answer a question.

REFERENCES

[1] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.
[2] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1821–1830.
[3] J. E. Haas, R. W. Kates, and M. J. Bowden, *Reconstruction Following Disaster*. Cambridge, MA, USA: Massachusetts Inst. Technol., 1977.
[4] S. Antol et al., "Visual question answering," in *Proc. ICCV*, 2015, pp. 1–4.
[5] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1564–1574.
[6] P. Gao et al., "Dynamic fusion with intra- and inter-modality attention flow for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6639–6648.
[7] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[8] J. Doshi, S. Basu, and G. Pang, "From satellite imagery to disaster insights," 2018, *arXiv:1812.07033*.

[9] T. G. Rudner et al., "Multi3Net: Segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 702–709.

[10] X. Zhu, J. Liang, and A. Hauptmann, "MSNet: A multilevel instance segmentation network for natural disaster damage assessment in aerial videos," 2020, *arXiv:2006.16479*.

[11] T. Chowdhury, M. Rahnemoonfar, R. Murphy, and O. Fernandes, "Comprehensive semantic segmentation on high resolution UAV imagery for natural disaster damage assessment," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2020, pp. 3904–3913.

[12] M. Turker and B. T. San, "Detection of collapsed buildings caused by the 1999 Izmit, Turkey earthquake through digital analysis of post-event aerial photographs," *Int. J. Remote Sens.*, vol. 25, no. 21, pp. 4701–4714, Nov. 2004, doi: 10.1080/01431160410001709976.

[13] S. A. Chen, A. Escay, C. Haberland, T. Schneider, V. Staneva, and Y. Choe, "Benchmark dataset for automatic damaged building detection from post-hurricane remotely sensed imagery," 2018, *arXiv:1812.05581*.

[14] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "FloodNet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89644–89654, 2021.

[15] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1–12.

[16] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 289–297.

[17] C. Kyrkou and T. Theocharides, "Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 517–525.

[18] A. Fujita, K. Sakurada, T. Imaizumi, R. Ito, S. Hikosaka, and R. Nakamura, "Damage detection from aerial images via convolutional neural networks," in *Proc. 15th IAPR Int. Conf. Mach. Vis. Appl. (MVA)*, May 2017, pp. 5–8.

[19] R. Gupta et al., "Creating XBD: A dataset for assessing building damage from satellite imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 10–17.

[20] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral Earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 2115–2118.

[21] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6172–6180.

[22] T. Feng et al., "Application and prospect of a high-resolution remote sensing and geo-information system in estimating earthquake casualties," *Natural Hazards Earth Syst. Sci.*, vol. 14, no. 8, pp. 2165–2178, Aug. 2014.

[23] J. F. Galarreta, N. Kerle, and M. Gerke, "UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning," *Natural Hazards Earth Syst. Sci.*, vol. 15, no. 6, pp. 1087–1101, Jun. 2015.

[24] T. Chowdhury and M. Rahnemoonfar, "Self attention based semantic segmentation on a natural disaster dataset," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2798–2802.

[25] T. Chowdhury, R. Murphy, and M. Rahnemoonfar, "RescueNet: A high resolution UAV semantic segmentation benchmark dataset for natural disaster damage assessment," 2022, *arXiv:2202.12361*.

[26] M. Rahnemoonfar, R. Murphy, M. V. Miquel, D. Dobbs, and A. Adams, "Flooded area detection from UAV images based on densely connected recurrent neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 1788–1791.

[27] F. Safavi and M. Rahnemoonfar, "Comparative study of real-time semantic segmentation networks in aerial images during flooding events," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 15–31, 2023.

[28] S. Tilon, F. Nex, N. Kerle, and G. Vosselman, "Post-disaster building damage detection from Earth observation imagery using unsupervised and transferable anomaly detecting generative adversarial networks," *Remote Sens.*, vol. 12, no. 24, p. 4193, Dec. 2020.

[29] J. Lee et al., "Assessing post-disaster damage from satellite imagery using semi-supervised learning techniques," 2020, *arXiv:2011.14004*.

[30] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

[31] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst. (CITS)*, Jul. 2016, pp. 1–5.

[32] G. Sumbul, S. Nayak, and B. Demir, "SD-RSIC: Summarization-driven deep remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6922–6934, Aug. 2021.

[33] S. Wu, X. Zhang, X. Wang, C. Li, and L. Jiao, "Scene attention mechanism for remote sensing image caption generation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.

[34] A. Sarkar and M. Rahnemoonfar, "UAV-VQG: Visual question generation framework on UAV images," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 4211–4219.

[35] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual question answering for remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8555–8566, Dec. 2020.

[36] S. Lobry, B. Demir, and D. Tuia, "RSVQA meets bigearthnet: A new, large-scale, visual question answering dataset for remote sensing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2021, pp. 1218–1221.

[37] C. Chappuis, S. Lobry, B. Kellenberger, B. Le Saux, and D. Tuia, "How to find a good image-text embedding for remote sensing visual question answering?" 2021, *arXiv:2109.11848*.

[38] X. Zheng, B. Wang, X. Du, and X. Lu, "Mutual attention inception network for remote sensing visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5606514.

[39] R. Felix, B. Repasky, S. Hodge, R. Zolfaghari, E. Abbasnejad, and J. Sherrah, "Cross-modal visual question answering for remote sensing data: The international conference on digital image computing: Techniques and applications (DICTA 2021)," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2021, pp. 1–9.

[40] A. Sarkar and M. Rahnemoonfar, "VQA-Aid: Visual question answering for post-disaster damage assessment and analysis," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2021, pp. 8660–8663.

[41] Z. Yuan, L. Mou, Z. Xiong, and X. X. Zhu, "Change detection meets visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5630613.

[42] C. T. Emrich, "Social vulnerability and hazard analysis for hurricane Harvey," Ph.D. thesis, School Public Admin., College Health Public Affairs Univ. Central Florida, Orlando, FL, USA, 2017.

[43] M. Ren, R. Kiros, and R. Zemel, "Image question answering: A visual semantic embedding model and a new dataset," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, vol. 1, no. 2, p. 5.

[44] J. Lu, X. Lin, D. Batra, and D. Parikh. (2015). *Deeper LSTM and Normalized CNN Visual Question Answering Model*. [Online]. Available: https://github.com/VT-vision-lab/VQA_LSTM_CNN

Argho Sarkar (Member, IEEE) received the B.S. degree in applied statistics from the University of Dhaka, Dhaka, Bangladesh, in 2018. He is currently pursuing the Ph.D. degree in information systems with the University of Maryland Baltimore County, Baltimore, MD, USA.

His research interests include on developing algorithms for multimodal applications such as visual question answering, and image captioning for medical and climate issues.

Tashnim Chowdhury (Member, IEEE) received the B.S. degree in electrical and electronic engineering from the Chittagong University of Engineering and Technology, Chittagong, Bangladesh, in 2013, and the M.S. degree in electrical engineering from The University of Toledo, Toledo, OH, USA, in 2016. He is currently pursuing the Ph.D. degree in information systems with the University of Maryland Baltimore County, Baltimore, MD, USA.

His research interests include deep learning, machine learning, semantic segmentation, few shot learning, meta learning, and bayesian learning.

Robin Roberson Murphy (Fellow, IEEE) received the B.M.E. degree in mechanical engineering, and the M.S. and Ph.D. degrees in computer science from Georgia Tech, Atlanta, Georgia, in 1980, 1989, and 1992, respectively.

She was a Rockwell International Doctoral Fellow at Georgia Tech. She is the Raytheon Professor of Computer Science and Engineering at Texas A&M University, College Station, TX, USA, and directs the Center for Robot-Assisted Search and Rescue. She is a Founder of the fields of rescue robots and human-robot interaction. She has over 100 publications including the best-selling textbook titled *Introduction to AI Robotics* (MIT Press, 2000). Her research interests are artificial intelligence, human-robot interaction, and heterogeneous teams of robots.

Aryya Gangopadhyay (Member, IEEE) received the Ph.D. degree in computer information systems from Rutgers University, Camden, NJ, USA.

He is a Professor with the Department of Information Systems, University of Maryland Baltimore County (UMBC), Baltimore, MD, USA. He has been a Faculty Member at UMBC since 1997. He has mentored and graduated 16 Ph.D. students who are working either as faculty members in various universities or holding leading IT positions in the private industries and government sectors. He has published five books and more than 125 peer-reviewed research articles. His research interests are in the area of data science and machine learning, machine learning-based solutions in areas such as cybersecurity, multimodal data fusion for emergency response, and healthcare applications such as computational drug repurposing.

Dr. Gangopadhyay research has been funded by grants from NSF, NIST, U.S. Department of Education, IBM, Maryland Department of Transportation, and other agencies. For more information, please visit https://sites.google.com/site/homearyya/.

Maryam Rahnemoonfar (Member, IEEE) received the Ph.D. degree in computer science from the University of Salford, Manchester, U.K., in 2010.

She is currently an Associate Professor and the Director of the Computer Vision and Remote Sensing Laboratory (Bina lab) at Lehigh University, Bethlehem, PA, USA. Her research interests include deep learning, computer vision, data science, AI for social good, remote sensing, and document image analysis. Her research specifically focuses on developing novel machine learning and computer vision algorithms for heterogeneous sensors such as radar, sonar, multi-spectral, and optical.

Dr. Rahnemoonfar's research has been funded by several awards including the NSF HDR Institute Award-iHARP, NSF BIGDATA Award, Amazon Academic Research Award, Amazon Machine Learning Award, Microsoft, and IBM.