IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 23, 2021, accepted June 2, 2021, date of publication June 21, 2021, date of current version June 29, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3090981

FloodNet: A High Resolution Aerial Imagery Dataset for Post Flood Scene Understanding

MARYAM RAHNEMOONFAR[1], (Member, IEEE), TASHNIM CHOWDHURY[1], (Graduate Student Member, IEEE), ARGHO SARKAR[1], DEBVRAT VARSHNEY[1], MASOUD YARI[1], AND ROBIN ROBERSON MURPHY[2], (Fellow, IEEE)

[1]Computer Vision and Remote Sensing Laboratory (Bina Lab), University of Maryland, Baltimore County, Baltimore, MD 21250, USA
[2]Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843, USA

Corresponding author: Maryam Rahnemoonfar (maryam@umbc.edu)

This work was supported in part by Microsoft and Amazon.

ABSTRACT Visual scene understanding is the core task in making any crucial decision in any computer vision system. Although popular computer vision datasets like Cityscapes, MS-COCO, PASCAL provide good benchmarks for several tasks (e.g. image classification, segmentation, object detection), these datasets are hardly suitable for post disaster damage assessments. On the other hand, existing natural disaster datasets include mainly satellite imagery which has low spatial resolution and a high revisit period. Therefore, they do not have a scope to provide quick and efficient damage assessment tasks. Unmanned Aerial Vehicle (UAV) can effortlessly access difficult places during any disaster and collect high resolution imagery that is required for aforementioned tasks of computer vision. To address these issues we present a high resolution UAV imagery, FloodNet, captured after the hurricane Harvey. This dataset demonstrates the post flooded damages of the affected areas. The images are labeled pixel-wise for semantic segmentation task and questions are produced for the task of visual question answering. FloodNet poses several challenges including detection of flooded roads and buildings and distinguishing between natural water and flooded water. With the advancement of deep learning algorithms, we can analyze the impact of any disaster which can make a precise understanding of the affected areas. In this paper, we compare and contrast the performances of baseline methods for image classification, semantic segmentation, and visual question answering on our dataset. FloodNet dataset can be downloaded from here: https://github.com/BinaLab/FloodNet-Supervised_v1.0.

INDEX TERMS Artificial intelligence, deep learning, hurricane Harvey, image classification, machine learning, natural disaster dataset, remote sensing, semantic segmentation, unmanned aerial vehicle (UAV), visual question answering.

I. INTRODUCTION

Visual scene understanding has the potential to advance many decision support systems. The purpose of scene understanding is to classify the overall category of a scene as well as to understand the interrelationship among different object classes at both instance and pixel level. Recently, several datasets [1]–[3] have been presented to study different aspects of scenes by implementing many computer vision tasks. A major factor in the success of most deep learning algorithms is the availability of large-scale datasets. Publicly available ground imagery datasets such as ImageNet [1], Microsoft COCO (Common Objects in Context) [2], PASCAL Visual Object Classes [3], Cityscapes [4] accelerate the advanced development of current deep neural networks, but aerial imagery data sets are scarce since the annotation is more tedious to obtain.

Aerial scene understanding datasets are helpful for urban management, city planning, infrastructure maintenance, damage assessment after natural disasters, and high definition maps for self-driving cars. Existing aerial datasets, however, are limited mainly to classification [5], [6] or semantic segmentation [5], [7] of few individual classes such as roads or buildings. Most of these datasets do not address the unique challenges in understanding post-disaster scenarios as a task for disaster damage assessment. Available post-disaster

The associate editor coordinating the review of this manuscript and approving it for publication was Halil Ersin Soken.

damage assessment datasets [8]–[11] mainly contain satellite images and images collected from social media. Satellite images are low in resolution and costly. On the other hand, images posted on social media are noisy and not scalable for deep learning models.

For quick response and recovery on large scale after a natural disaster such as a hurricane, wildfire, and extreme flooding access to high-resolution aerial images are critically important for the response team. To fill this gap we present *FloodNet* dataset associated with three different computer vision tasks namely classification, semantic segmentation, and visual question answering (VQA).

FloodNet, provides high-resolution images taken from low altitude, as compared to satellite images which capture images from a higher altitude and may have obstructions from clouds and smoke. The characteristics of *FloodNet* bring more clarity to scenes and can help deep learning models in making more accurate decisions regarding post-disaster damage assessment. Currently, most of the tasks considering natural disaster datasets are restricted to mainly classification and object detection. Our dataset offers pixel-level annotation for semantic segmentation and VQA, besides image classification. All these three computer vision tasks can assist in the understanding of a scene and help rescue teams efficiently manage their operations during emergencies. Figure 1 shows sample annotations offered by FloodNet.

Our contribution is two folds. First, we introduce high-resolution UAV imagery with pixel-level annotations named *FloodNet* for post-disaster damage assessment. Secondly, we compare the performance of several classification, semantic segmentation, and VQA methods on our dataset. To the best of our knowledge, this is the first semantic segmentation and VQA work focused on UAV imagery for any disaster damage assessment.

The remainder of this paper is organized as follows - it begins with highlighting the existing datasets for natural disasters and also describes the computer vision tasks of image classification, semantic segmentation, and VQA in section II. Next, section III describes the *FloodNet* dataset including its collection and annotation process. Section IV describes the experimental setups for all three aforementioned tasks and section V gives a complete analysis of the results. Finally section VI summarizes the results including conclusion and future works.

II. RELATED WORKS

In this section we provide an overview of datasets designed for natural disasters damage analysis, followed by a survey of techniques targeting aerial and satellite image classification, segmentation, and VQA.

A. DATASETS

Natural disaster datasets are of two types: A) non-imaging dataset (text, tweets, social media posts) [28], [29] and B) imaging dataset [5], [7], [24]. Based on the position of the captured image, current image-based natural disaster datasets can be classified into three classes: B1) ground-level images [30], B2) satellite imagery [5], [7], [23]–[26], and B3) aerial imagery [6], [22], [27]. Recently several datasets have been introduced by researchers for natural disaster damage assessment. Nguyen *et al.* proposed an extension of AIDR system [21] to collect data from social media in [30]. AIST Building Change Detection (ABCD) dataset has been proposed in [22] which includes aerial post tsunami images to identify whether the buildings have been washed away. A combination of SpaceNet [31] and DeepGlobe [32] was presented in [23] and a segmentation model was proposed to detect changes in man-made structures and estimate the impact of natural disasters. Chen *et al.* in [24] proposed a fusion of different data resources for automatic building damage detection after a hurricane. The dataset includes satellite and aerial imageries along with vector data. Onera Satellite Change Detection (OSCD) dataset was proposed in [25] which consists of multispectral aerial images to detect urban growth and changes with time. A collection of images of buildings and lands named Functional Map of the World (fMoW) was introduced by Christie *et al.* in [26]. Aerial Image Database for Emergency Response (AIDER) is proposed by Kyrkou and Theocharides in [6] for classification of UAV imagey. Rudner *et al.* [7] proposed a satellite imagery dataset collected from Sentinel-1 and Sentinel-2 satellites for semantic segmentation of flooded buildings. Gupta *et al.* proposed xBD [5] which have both pre- and post-event satellite images in order to assess building damages. Recently ISBDA (Instance Segmentation in Building Damage Assessment) is created by Zhu *et al.* in [27] for instance segmentation while images are collected using UAVs.

A comparative study among different disaster and non-disaster datasets is shown in Table 1. As you can see in Table 1, our dataset is the only high resolution UAV dataset collected after a hurricane which contains all computer vision tasks including classification, semantic segmentation, and VQA. Although several pre- and post-disaster datasets have been proposed over the years, these datasets primarily consist of satellite images. Satellite imageries, including those with high resolution, do not provide enough details about the post disaster scenes which are necessary to distinguish among different damage categories of different objects. On the other hand the primary source of the ground-level imageries is social media [30], these imageries lack geo-location tags [27] and suffers from data scarcity for deep learning training [11]. Although some aerial datasets [6], [27] are prepared using UAVs, these datasets lack low altitude high resolution images. AIDER [6] dataset collected images from different sources for image classification task and contains far more examples of normal cases rather than damaged objects; therefore lacks consistency and generalization. ISBDA [27] provides only building instance detection capability rather than inclusion of other damaged objects and computer vision tasks like semantic segmentation and VQA. To address all these issues, FloodNet includes low altitude high resolution post disaster

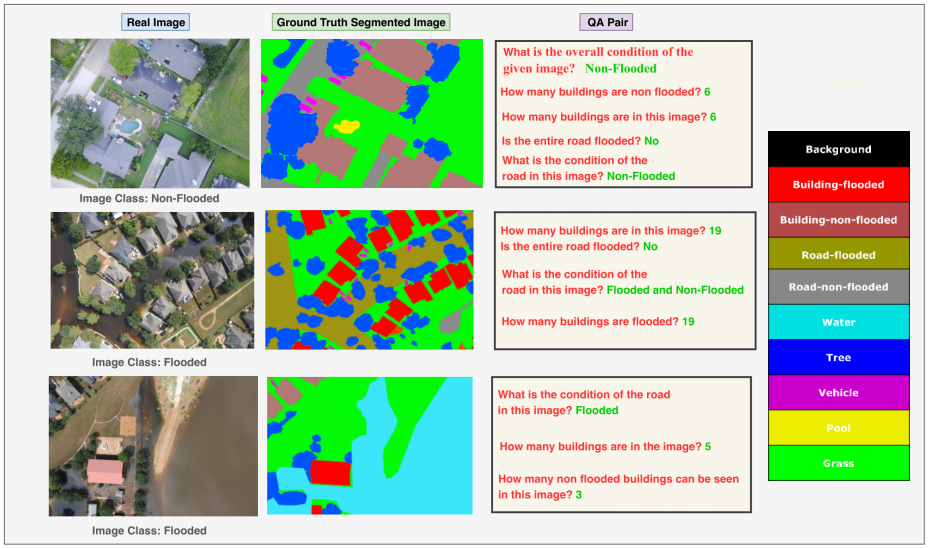

FIGURE 1. FloodNet dataset overview for Classification, Semantic Segmentation and Visual Question Answering.

TABLE 1. A brief summary of existing datasets.

Dataset	Types of Images	UAV imagery	Post Disaster	Resolution of Images	Classification	Semantic Segmentation	VQA
ImageNet [1]	Real-world images	No	No	average 400 × 350	✓	✗	✗
Cityscapes [2]	Real-world images	No	No	1280 × 720	✗	✓	✗
DAQUAR [12]	Real-world images	No	No	640 × 480	✗	✗	✓
COCO-QA [13]	Real-world images	No	No	640 × 480	✗	✗	✓
COCO-VQA [14]	Real world images, abstract cartoon images	No	No	640 × 480	✗	✗	✓
Visual Genome [15]	Real-world images	No	No	varies in size	✗	✗	✓
Visual7W [16]	Real-world images	No	No	varies in size	✗	✗	✓
TDIUC [17]	Real-world images	No	No	varies in size	✗	✗	✓
CLEVR [18]	Geometrical Shape	No	No	320 x 240 (in default settings)	✗	✗	✓
PATHVQA [19]	Medical Images	No	No	Varies in size	✗	✗	✓
VQA-MED [20]	Medical Images	No	No	Varies in size	✗	✗	✓
Nguyen et al. [21]	Post Disaster Images	No	Yes	Varies in size	✓	✗	✗
ABCD [22]	Pre and Post Disaster Images	No	Yes	Varies in size	✓	✗	✗
SpaceNet + Deepglobe [23]	Pre and Post Disaster Images	No	Yes	Varies in size	✗	✓	✗
Chen et al. [24]	Post Disaster Images	No	Yes	Varies in size	✗	✗	✗
OSCD [25]	Urban Change Images	No	No	Varies in size	✗	✗	✗
fMoW [26]	Pre and Post Disaster Images	No	Yes	Varies in size	✓	✗	✗
AIDER [6]	Post Disaster Images	Yes	Yes	Varies in size	✓	✗	✗
Rudner et al. [7]	Post Disaster Images	No	Yes	Varies in size	✗	✓	✗
xBD [5]	Pre and Post Disaster Images	No	Yes	Varies in size	✓	✓	✗
ISBDA [27]	Post Disaster Images	Yes	Yes	Varies in size	✗	✗	✗
FloodNet (Ours)	**Post Disaster Images**	**Yes**	**Yes**	**4000× 3000**	✓	✓	✓

images annotated for classification, semantic segmentation, and VQA. FloodNet provides more details about the scenarios which help to estimate the post disaster damage assessment more accurately.

B. ALGORITHMS

Here we review the related computer vision algorithms and some of their applications in disaster damage assessment.

1) CLASSIFICATION

The utility of deep neural networks was realized when they achieved high accuracy in categorizing images into different classes. This was given a boost mainly by Krizhevsky *et al.* [33] which achieved state-of-the-art performance on the ImageNet [1] dataset in 2012. As this is arguably the most primitive computer vision task, a lot of networks were proposed subsequently which could perform classification on public datasets such as CIFAR [34], [35], MNIST [36], and FashionMNIST [37].

This led to a rise in networks such as VGGNet [38], ResNet [39], InceptionNet [40], Xception [41], MobileNet [42] etc., where the network architectures were experimented with different skip connections, residual learning, multi-level feature extraction, separable convolutions, and optimization methods for mobile devices. Although these networks achieved good performance on day to day images of animals and vehicles, they were hardly sufficient to make predictions on scientific datasets such as those captured by air-borne or space-borne sensors.

In this regard, some image classification networks have been explored for the purpose of post-disaster damage detection, such as [21], [43]–[46]. [21] used crowd sourced images from social media which captured disaster sites from the ground level. [44] used a Support Vector Machine on top of a

Convolutional Neural Network (CNN) followed by a Hidden Markov Model post-processing to detect avalanches. [45] compared [38] and [39] for fire detection, but then again the dataset used contained images taken by hand-held cameras on the ground. [46] developed a novel algorithm which focused on wildfire detection through UAV images. [43] have done extensive work by developing a CNN for emergency response towards fire, flood, collapsed buildings, and crashed cars.

2) SEMANTIC SEGMENTATION

Semantic segmentation is one of the prime research areas in computer vision and an essential part of scene understanding. Fully Convolutional Network (FCN) [47] is a pioneering work which is followed by several state-of-art models to address semantic segmentation. From the perspective of contextual aggregation, segmentation models can be divided into two types. Models, such as PSPNet (Pyramid Scene Parsing Network) [48] or DeepLab [49], [50] perform spatial pyramid pooling [51], [52] at several grid scales and have shown promising results on several segmentation benchmarks. The encoder-decoder networks combines mid-level and high-level features to obtain global context from different scales. Some notable works using this architecture are [50], [53]. On the other hand, there are models [54]–[56] which obtain feature representation by learning contextual dependencies over local features.

Besides proposing natural disaster datasets many researchers have also presented different deep learning models for post natural disaster damage assessment. Authors in [23] perform previously proposed semantic segmentation [57] on satellite images to detect changes in the structure of various man-made features, and thus detect areas of maximal impact due to natural disaster. Rahnemoonfar *et al.* present a densely connected recurrent neural network in [58] to perform semantic segmentation on UAV images for flooded area detection. Rudner *et al.* fuse multiresolution, multisensor, and multitemporal satellite imagery and propose a novel approach named Multi3Net in [7] for rapid segmentation of flooded buildings. Gupta *et al.* propose a DeepLabv3 [50] and DeepLabv3+ [59] inspired RescueNet in [60] for joint building segmentation and damage classification. All these proposed methods address the semantic segmentation of specific object classes like river, buildings, and roads rather than complete scene post disaster scenes.

Above mentioned state-of-art semantic segmentation models have been primarily applied on ground based imagery [4], [61]. In contrast we apply three state-of-art semantic segmentation networks on our proposed FloodNet dataset. We adopt one encoder-decoder based network named ENet [62], one pyramid pooling module based network PSPNet [48], and the last model DeepLabv3+ [59] employs both encoder-decoder and pyramid pooling based modules.

3) VISUAL QUESTION ANSWERING (VQA)

Many researchers proposed several datasets and methods for VQA task.

To find the right answer, VQA systems need to model the question and image (visual content). Substantial research efforts have been made on the VQA task based on real natural and medical imagery in the computer vision and natural language processing communities [14], [63]–[65] using deep learning-based multimodal methods [66]–[74]. In these methods, different approaches for the fined-grained fusion between semantic features of image and question have been proposed. Most of the recent VQA algorithms have trained on natural image based datasets such as DAQUAR (Dataset for Question Answering on Real-world images) [75], COCO-VQA [14], Visual Genome [15], Visual7W [16]. In addition Path-VQA [19] and VQA-MED [20] are medical images for which VQA algorithms are also considered. There are no such datasets apt for training and evaluating VQA algorithms regarding disaster damage assessment task. In this work, we present *FloodNet* dataset to build and test VQA algorithms that can be implemented during natural emergencies. To the best of our knowledge, this is the first VQA dataset focused on UAV imagery for disaster damage assessment. To evaluate the performances of existing VQA algorithms we have implemented baseline models, Stacked Attention network [63], and MFB with Co-Attention [74] network on our dataset.

III. THE FloodNet DATASET

The data is collected with small UAV platform, DJI Mavic Pro quadcopters, after *Hurricane Harvey*. Hurricane Harvey made landfall near Texas and Louisiana on August, 2017, as a Category 4 hurricane. The Harvey dataset consists of video and imagery taken from several flights conducted between August 30 - September 04, 2017, at Ford Bend County in Texas and other directly impacted areas. The dataset is unique for two reasons. One is fidelity: it contains imagery from sUAV taken during the response phase by emergency responders, thus the data reflects what is the state of the practice and can be reasonable expected to be collected during a disaster. Second: it is the only known database of sUAV imagery for disasters. Note that there are other existing databases of imagery from unmanned and manned aerial assets collected during disasters, such as National Guard Predators or Civil Air Patrol, but those are larger, fixed-wing assets that operate above the 400 feet AGL (Above Ground Level), limitation of sUAV. All flights were flown at 200 feet AGL, as compared to manned assets which normally fly at 500 feet AGL or higher. At a height of 200 feet, our images correspond to a very high spatial resolution, about 1.5cm, making them unique compared to other datasets for natural disasters. The post-flooded damages to affected areas are demonstrated in all the images. There are several objects (e.g. construction, road) and related attributes (e.g. state of an object such as flooded or non-flooded after Hurricane Harvey) represented by these images. For the preparation of this dataset for semantic segmentation and VQA, these attributes are considered. FloodNet dataset can be downloaded from here: https://github.com/BinaLab/FloodNet-Supervised_v1.0

A. ANNOTATION TASKS

After natural disasters, the response team first need to identify the affected neighborhoods such as flooded neighborhoods (classification tasks). Then on each neighborhood they need to identify flooded buildings and roads (semantic segmentation) so the rescue team can be sent to affected areas. Furthermore, damage assessment after any natural calamities done by querying about the changes in object's condition so they can allocate the right resources. Based on these needs and with the help of response and rescue team, we defined classification, semantic segmentation and VQA tasks.

In total 2343 images have been annotated with 9 classes which include building-flooded, building-non-flooded, road-flooded, road-non-flooded, water, tree, vehicle, pool, and grass. A buildings is classified as flooded when at least one side of a building is touching the flood water. Although we have classes created for flooded buildings and roads, to distinguish between natural water and flood water, ''water'' class has been created which represents any natural water body like river and lake. For the classification task, each image is classified either ''flooded'' or ''non-flooded''. If more than 30% area of an image is occupied by flood water then that area is classified as flooded, otherwise non-flooded. Number of images and instances corresponding to different classes are shown in Table 2. Our images are quite dense. On average, it takes about one hour to annotate each image. To ensure high quality, we performed the annotation process iteratively with a two-level quality check over each class. The images are annotated on V7 Darwin platform [76] for classification and semantic segmentation. Annotation tasks on V7 Darwin platform is performed in two steps. In the first step the image is assigned to an annotator randomly. After the annotation is complete, the images are sent to the reviewers. Depending on the quality of the annotation, the images are either being accepted or sent back to the annotators with comments. The review and feedback cycle continues until the annotation reaches the high quality.

We split the dataset into training, validation, and test sets with 70% for training and 30% for validation and testing. The training, validation, and testing sets for all the three tasks will be publicly available.

TABLE 2. Number of images and instances corresponding to different classes.

Object Class	Images	Instances
Building-flooded	245	3248
Building-non-flooded	880	3427
Road-flooded	264	495
Road-non-flooded	1175	2155
Vehicle	813	4535
Pool	531	1141
Tree	1885	19682
Water	984	1374

B. VQA TASK

To provide VQA framework, we focus on generating questions related to the building, road, and entire image as a whole for our *FloodNet* dataset. By asking questions related to these objects we can assess the damages and understand the situation very precisely. Attribute associated with aforementioned objects can be identified from the Table 2. For the FloodNet-VQA dataset, $\sim$ 4500 question-image pairs are considered while training VQA networks. All the questions are created manually. Each image has an average of 3.5 questions. Each of the questions is designed to provide answers which are connected to the local and global regions of images. In Figure 1, some sample questions-answer pairs are presented from our dataset.

1) TYPES OF QUESTION

Questions are divided into a four-way question group, namely *"Simple Counting"*, *"Complex Counting"*, *"yes/no"*, and *"Condition Recognition"*. In the Figure 2, distribution of the question pattern based on the first words of the questions is given. All of the questions start with a word belonging to the set {How, Is, What}. Maximum length of question is 11.

In the *Simple Counting* problem, we ask about an object's frequency of presence (mainly building) in an image, regardless of the attribute (e.g. *How many buildings are in the images?*). Both flooded and non-flooded buildings can appear in a picture in several cases (e.g. bottom image from Figure 1).

The question type *Complex Counting* is specifically intended to count the number of a particular building attribute (e.g. *How many* ***flooded/non-flooded*** *buildings are in the images?*) We're interested in counting only the flooded or non-flooded buildings from this type of query. In comparison to simple counting, a high-level understanding of the the scene is important for answering this type of question. This type of question also starts with the word ''How''.

Condition Recognition questions investigate the condition of the entire image as a whole or any object. ''What is the condition of the road?'', ''What is the overall condition of the entire image?'' are the examples from this category, Starting word for this type of question is ''What''.

Yes/No type question is categorised as the fourth type of question. ''Is the road flooded?'', ''Is the road non-flooded?'' are some of the examples from this category. Starting word for this type of question is ''Is''.

2) TYPES OF ANSWER

Both flooded and non-flooded buildings can exist in any image. For complex counting problem, we only count either the flooded or non-flooded buildings from a given image-question pair. Roads are also annotated as flooded or non-flooded. Second image from the Figure 1 depicts both flooded and non-flooded roads. Thus, the answer for the question like ''What is condition of road?'' for this kind of images will be both 'flooded and non-flooded'. Furthermore, entire image may be graded as flooded or non-flooded. Table 3 refers to the possible answers for three types questions and from Figure 2, we can see the possible answer distribution for different types of question. Most frequent answers for counting problem, in general, are '4, 3, 2, 1' whereas '27,

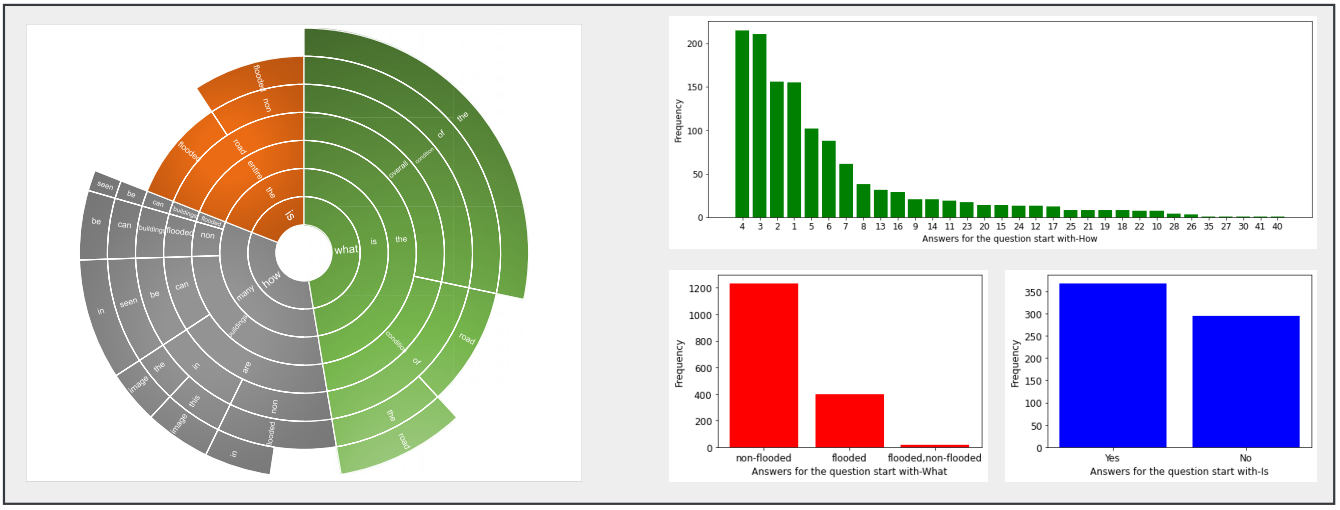

FIGURE 2. VQA Data Statistics: Left figure represents the distribution of the question pattern based on starting word; Right-top, Right-bottom figures describes the distribution of possible answers for different question types.

TABLE 3. Set of possible answers for three types of questions.

Question Type	Possible Answer
Simple Counting	{1,2,3,4...}
Complex Counting	{1,2,3,4...}
Condition Recognition	Flooded , Non-Flooded, Both Flooded & Non-Flooded
Yes/No-Type Question	Yes, No

30, 41, 40' are the less frequent answers. For *Condition Recognition* problem, 'non-flooded', and for *yes/no* type of question, 'yes' is the most common answer.

IV. IMPLEMENTATION

To understand the usability of these images for flood detection, we majorly carry out three tasks, which are image classification, semantic segmentation, and VQA. In this section we describe how we implement the three tasks, and the hyperparameters involved. For all of our tasks, we use NVIDIA GeForce RTX 2080 Ti GPU with an Intel Core i9 processor.

For network evaluation, we use mean intersection over union (mIoU) and accuracy as the metrics defined in Equations 1 and 2 respectively. In these equations k represents the total number of classes, whereas TP, TN, FP, FN represent true positives, true negatives, false positives, and false negatives respectively.

$$\text{mIoU} = \frac{1}{k}\sum_{i=1}^{k}\frac{TP_i}{TP_i + FP_i + FN_i} \tag{1}$$

$$\text{Accuracy} = \frac{1}{k}\sum_{i=1}^{k}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{2}$$

A. IMPLEMENTATION OF IMAGE CLASSIFICATION

For image classification, we used three state-of-the-art networks i.e. InceptionNetv3 [77], ResNet50 [39], and Xception [41] as base models to classify the images into Flooded and Non-Flooded categories. Each of these networks have significantly contributed to the field of computer vision by introducing a unique design element, such as the residual blocks in ResNet, the multi-scale architecture in InceptionNet and depthwise separable convolutions in Xception. For our classification task, the output from these base models was followed by a global average pooling layer, a fully connected layer with 1024 neurons having rectified linear unit (RELU) activation, and finally by two neurons with softmax activation. We initialized our networks with ImageNet [1] weights and trained them for 30 epochs, with 20 steps for every epoch, using binary cross entropy loss. We used the Adam optimizer [78] with a learning rate of 0.001, and resized all images to 224 × 224 dimension.

B. IMPLEMENTATION OF SEMANTIC SEGMENTATION

For semantic segmentation, we implemented three methods, i.e. PSPNet [48], ENet [62], and DeepLabv3+ [59]; and evaluate their performance on FloodNet dataset. For implementing PSPNet, ResNet101 was used as backbone. We used ''poly'' learning rate with base learning rate 0.0001. Momentum, weight decay, power, and weight of the auxiliary loss were set to 0.9, 0.0001, 0.9, and 0.4 respectively. For ENet we used 0.0005 and 0.1 for learning rate and learning rate decay respectively. Weight decay was set to 0.0002. Similarly for DeepLabv3+ we used poly learning rate with base learning rate 0.01. We set weight decay to 0.0001 and momentum to 0.9. For image augmentation we used random shuffling, scaling, flipping, and random rotation which helped the models avoid overfitting. From different experiments it was proved that larger ''crop size'' and ''batch size'' improve the performance of the models. During training, we resized the images to 713 × 713 since large crop size is useful for the high resolution images. For semantic segmentation evaluation metric, we used mean intersection over union (mIoU).

TABLE 4. Per-class intersection over union (in %) and their mean value (mIoU) on FloodNet testing set.

Method	Building Flooded	Building Non Flooded	Road Flooded	Road Non Flooded	Water	Tree	Vehicle	Pool	Grass	mIoU
ENet [62]	21.82	41.41	14.76	52.53	47.14	62.56	26.21	16.57	75.57	39.84
DeepLabv3+ [59]	28.10	78.10	32.00	81.10	73.00	74.50	33.60	40.00	87.10	58.61
PSPNet [48]	**65.61**	**90.92**	**78.69**	**90.90**	**91.25**	**89.17**	**54.83**	**66.37**	**95.45**	**80.35**

C. IMPLEMENTATION OF VQA

Finally, for VQA, simple baselines (concatenation/element-wise product of image and text features) and Multimodal Factorized Bilinear (MFB) with co-attention [74], Stacked Attention Network [63] have been considered for this study. All these models are configured according to our dataset. We do not considered any pre-trained weights for image feature extraction. For image and question feature extraction, respectively, VGGNet (VGG 16) [38] and Two-Layer LSTM (Long short-term memory) [80] are taken into account. Feature vector from last pooling layer of the VGGNet and 1024-D vector from the last layer of Two-Layer LSTM are extracted as the image and question vectors respectively. Dataset is split into training, validation and testing data. All the images are resized to 224 $\times$ 224 and questions are tokenized before feed into the model. All the questions are converted into lower-case and punctuation are removed in the text pre-processing step. By considering cross-entropy loss, all the models are optimized by stochastic gradient descent (SGD) with batch size 16. In the training phase, models are validated by validation dataset via early stopping criterion with patience 30.

V. RESULTS

We have implemented baseline models on our *FloodNet* dataset for three computer vision tasks namely image classification, semantic segmentation and VQA. In this section we will present the results from baseline models for all the three tasks individually.

A. IMAGE CLASSIFICATION PERFORMANCE ANALYSIS

The classification accuracies of the three networks are shown in Table 5. From this table, it can be seen that although all three networks give similar performance, the highest performance on the test set was given by InceptionNetv3. The multi-scale architecture of this network has successfully helped in classifying the test images into Flooded and Non-Flooded classes, as compared to other networks. The depthwise separable convolutions of Xception and residual architecture of ResNet50 gave slightly worse performance, although Xception showed the highest performance on the training set. This is contrary to the networks' performance on the ImageNet dataset where ResNet50 gave the highest performance.

Therefore, networks which give high accuracy on everyday images such as those of ImageNet cannot really be used to detect image features from aerial datasets which contain more complex urban and natural scenes. Thus, there is a need to design separate novel architectures which can effectively detect urban disasters.

TABLE 5. Classification accuracy (in %) of three state-of-the-art networks on the training, validation, and test sets of FloodNet data.

Model	Training	Validation	Test
InceptionNetv3 [77]	95.37	92.89	95.09
ResNet50 [39]	93.99	92.22	93.30
Xception [41]	96.54	96.00	94.64

B. SEMANTIC SEGMENTATION PERFORMANCE ANALYSIS

Semantic segmentation results of ENet, DeepLabv3+, and PSPNet are presented in Table 4. From the segmentation experiment it is evident that detecting small objects like vehicles and pools are the most difficult tasks for the segmentation networks. Flooded buildings and roads are the next challenging tasks for all three models. Among all of the segmentation models, PSPNet performs best in all classes. It is interesting to note that although DeepLabv3+ and PSPNet collect global contextual information, their performances on detecting flooded buildings and flooded roads are still low, since distinguishing between flooded and non-flooded objects heavily depends on respective contexts of the classes. The qualitative results of these three networks are shown in Figure 3.

C. VQA PERFORMANCE ANALYSIS

Accuracy is the performance metric that we consider for the VQA task to compare the baseline models. We consider top-1 accuracy for the comparison purpose. If the ground-truth matches the output (which has the highest probability) from a model, the accuracy for any image is 1, otherwise it is 0. From the Table 6, we can identify that counting problem (simple and complex) is very challenging compared to task of condition recognition. Many objects are very small which makes it very difficult even for humans to count. Accuracy for *'Condition Recognition'* category is consistent for all of the methods. This is because it is not difficult to recognize the condition of whole images as well as roads as they are pictured in a larger ratio given the overall size of an image. Performances of the all methods for *'Condition Recognition'* category are when we compare those with other categories. MFB with co-attention [74] performs well for 'yes/no' type of question. Stacked Attention Network [63] shows better result for all counting (e.g. simple and complex) related problem compare to the other methods.

TABLE 6. Comparison of accuracy between baseline VQA algorithms on our dataset.

Mode of Feature Combination/Model	Data Type	Overall Accuracy	Accuracy for 'Simple Counting'	Accuracy for 'Complex Counting'	Accuracy for 'Yes/No'	Accuracy for 'Condition Recognition'
Concatenation [79]	Validation	0.53	0.06	0.05	0.33	0.88
	Testing	0.52	0.06	0.03	0.31	0.88
Point-wise Multiplication [14]	Validation	0.77	0.32	0.29	0.95	0.96
	Testing	0.77	0.3	0.28	0.97	**0.97**
MFB with Co-Attention [74]	Validation	0.77	0.27	0.26	0.98	0.97
	Testing	0.74	0.21	0.21	**0.98**	0.96
SAN [63]	Validation	0.7	0.31	0.30	0.58	0.97
	Testing	0.71	**0.32**	**0.32**	0.62	0.96

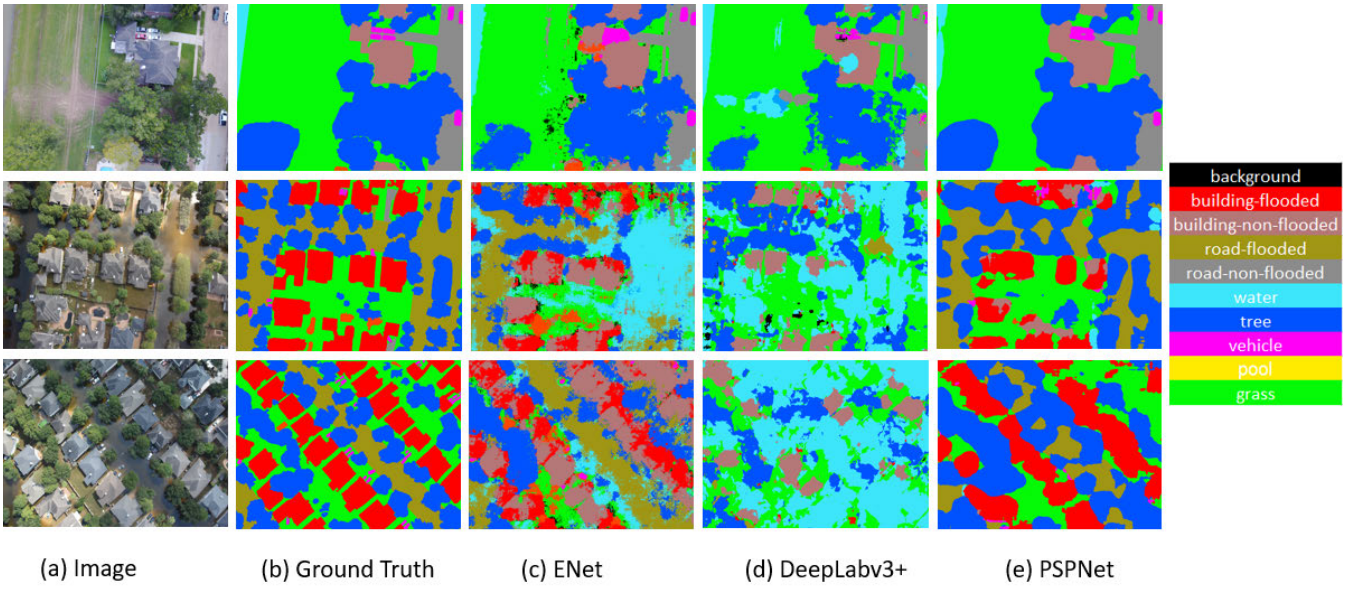

FIGURE 3. Visual comparison on FloodNet test set for Semantic Segmentation.

VI. DISCUSSION AND CONCLUSION

In this paper, we introduce the FloodNet dataset for post natural disaster damage assessment. We describe the dataset collection procedure along with different features and statistics. The UAV images provide high resolution and low altitude dataset specially significant for performing computer vision tasks. The dataset is annotated for classification, semantic segmentation, and VQA. We perform three computer vision tasks including image classification, semantic segmentation, and VQA and in-depth analysis have been provided for all three tasks.

Although UAVs are cost effective and prompt solution during any post natural disaster damage assessment, several challenges have been posed by FloodNet dataset collected using UAVs. Among all the existing classes, vehicles and pools are the smallest in shape and therefore would be difficult for any network models to detect them. Segmentation results from Table 4 supports the task difficulty in identifying small objects like vehicles and pools. Besides detecting flooded building is another prime challenge. Since UAV images only include top view of a building, it is very difficult to estimate how much damages are done on that building. Segmentation models do not perform well in detecting flooded buildings. Similarly flooded roads pose challenge in distinguishing them from non-flooded roads and results from segmentation models prove that. Most importantly distinguishing between flooded and non-flooded roads and buildings depends on their corresponding contexts and current state-of-art models are still lacking good performance in computer vision tasks performed on FloodNet. To the best of our knowledge this is the first time where these three crucial computer vision tasks have been addressed in a post natural disaster dataset together. The experiments of the dataset show great challenges and we strongly hope that FloodNet will motivate and support the development of more sophisticated models for deeper semantic understanding and post disaster damage assessment.

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[2] T.-Y. Lin, M. Maine, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," 2015, *arXiv:1405.0312*. [Online]. Available: https://arxiv.org/abs/1405.0312

[3] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[5] R. Gupta, B. Goodman, N. Patel, R. Hosfelt, S. Sajeev, E. Heim, J. Doshi, K. Lucas, H. Choset, and M. Gaston, "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.
[6] 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.
[7] T. G. Rudner, M. Rußwurm, J. Fil, R. Pelich, B. Bischke, V. Kopačková, and P. Biliński, "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.
[8] B. Bischke, P. Helber, C. Schulze, V. Srinivasan, A. Dengel, and D. Borth, "The multimedia satellite task at mediaeval 2017," in *Proc. MediaEval*, 2017.
[9] B. Benjamin, H. Patrick, Z. Zhengyu, and B. Damian, "The multimedia satellite task at mediaeval 2018: Emergency response for flooding events," Tech. Rep., 2018.
[10] R. Gupta, R. Hosfelt, S. Sajeev, N. Patel, B. Goodman, J. Doshi, E. Heim, H. Choset, and M. Gaston, "XBD: A dataset for assessing building damage from satellite imagery," 2019, *arXiv:1911.09296*. [Online]. Available: http://arxiv.org/abs/1911.09296
[11] E. Weber, N. Marzo, P. Dim Papadopoulos, A. Biswas, A. Lapedriza, F. Ofli, M. Imran, and A. Torralba, "Detecting natural disasters, damage, and incidents in the wild," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 331–350.
[12] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1682–1690.
[13] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2953–2961.
[14] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Visual question answering," in *Proc. ICCV*, 2015, pp. 2425–2433.
[15] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual Genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.
[16] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded question answering in images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4995–5004.
[17] K. Kafle and C. Kanan, "An analysis of visual question answering algorithms," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1965–1973.
[18] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2901–2910.
[19] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie, "PathVQA: 30000+ questions for medical visual question answering," 2020, *arXiv:2003.10286*. [Online]. Available: http://arxiv.org/abs/2003.10286
[20] A. B. Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, and H. Müller, "VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019," in *Proc. CLEF Work. Notes*, 2019.
[21] D. Tien Nguyen, F. Alam, F. Ofli, and M. Imran, "Automatic image filtering on social networks using deep learning and perceptual hashing during crises," 2017, *arXiv:1704.02602*. [Online]. Available: http://arxiv.org/abs/1704.02602
[22] 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.
[23] J. Doshi, S. Basu, and G. Pang, "From satellite imagery to disaster insights," 2018, *arXiv:1812.07033*. [Online]. Available: http://arxiv.org/abs/1812.07033
[24] 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*. [Online]. Available: http://arxiv.org/abs/1812.05581
[25] R. C. Daudt, B. L. 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. (IGARSS)*, Jul. 2018, pp. 2115–2118.
[26] 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.
[27] 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*. [Online]. Available: http://arxiv.org/abs/2006.16479
[28] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: A survey," *ACM Comput. Surveys*, vol. 47, no. 4, pp. 1–38, Jul. 2015.
[29] C. Reuter and M.-A. Kaufhold, "Fifteen years of social media in emergencies: A retrospective review and future directions for crisis informatics," *J. Contingencies Crisis Manage.*, vol. 26, no. 1, pp. 41–57, Mar. 2018.
[30] D. T. Nguyen, F. Ofli, M. Imran, and P. Mitra, "Damage assessment from social media imagery data during disasters," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Jul. 2017, pp. 569–576.
[31] *NVIDIA: Spacenet on Amazon Web Services (AWS) Datasets: The Spacenet Catalog*, DigitalGlobe CosmiQWorks, 2016.
[32] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "DeepGlobe 2018: A challenge to parse the earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.
[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[34] A. Krizhevsky, V. Nair, and G. Hinton, "CIFAR-10 (canadian institute for advanced research)," Tech. Rep., 2009.
[35] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.
[36] Y. LeCun, C. Cortes, and C. Burges. (Feb. 2010). *Mnist Handwritten Digit Database*. ATT Labs. [Online]. Available: http://yann.lecun.com/exdb/mnist
[37] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*. [Online]. Available: http://arxiv.org/abs/1708.07747
[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[41] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.
[42] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861
[43] C. Kyrkou and T. Theocharides, "EmergencyNet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1687–1699, 2020.
[44] M. Bejiga, A. Zeggada, A. Nouffidj, and F. Melgani, "A convolutional neural network approach for assisting avalanche search and rescue operations with UAV imagery," *Remote Sens.*, vol. 9, no. 2, p. 100, Jan. 2017.
[45] J. Sharma, O.-C. Granmo, M. Goodwin, and J. T. Fidje, "Deep convolutional neural networks for fire detection in images," in *Proc. Int. Conf. Eng. Appl. Neural Netw.* New York, NY, USA: Springer, 2017, pp. 183–193.
[46] Y. Zhao, J. Ma, X. Li, and J. Zhang, "Saliency detection and deep learning-based wildfire identification in UAV imagery," *Sensors*, vol. 18, no. 3, p. 712, Feb. 2018.
[47] 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.
[48] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[49] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[50] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587
[51] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 1458–1465.
[52] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 2169–2178.
[53] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2015, pp. 234–241.
[54] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 267–283.
[55] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.
[56] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[57] J. Doshi, "Residual inception skip network for binary segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 216–219.
[58] 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. (IGARSS)*, Jul. 2018, pp. 1788–1791.
[59] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
[60] R. Gupta and M. Shah, "RescueNet: Joint building segmentation and damage assessment from satellite imagery," 2020, *arXiv:2004.07312*. [Online]. Available: http://arxiv.org/abs/2004.07312
[61] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 891–898.
[62] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*. [Online]. Available: http://arxiv.org/abs/1606.02147
[63] 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.
[64] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1564–1574.
[65] P. Gao, Z. Jiang, H. You, P. Lu, S. C. H. Hoi, X. Wang, and H. Li, "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.
[66] M. I. Hasan Chowdhury, K. Nguyen, S. Sridharan, and C. Fookes, "Hierarchical relational attention for video question answering," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 599–603.
[67] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proc. ECCV*. Cham, Switzerland: Springer, 2016, pp. 451–466.
[68] 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. (EMNLP)*, 2016, pp. 457–468.
[69] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "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.
[70] D. Yu, J. Fu, X. Tian, and T. Mei, "Multi-source multi-level attention networks for visual question answering," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 2s, pp. 1–20, Aug. 2019.
[71] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 5947–5959, Dec. 2018.
[72] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2612–2620.
[73] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," 2016, *arXiv:1610.04325*. [Online]. Available: http://arxiv.org/abs/1610.04325
[74] 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.
[75] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Computer Vision—ECCV*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Germany: Springer, 2012, pp. 746–760.
[76] *V7 Darwin*. Accessed: Nov. 11, 2020. [Online]. Available: https://www.v7labs.com/darwin
[77] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[78] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[79] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," 2015, *arXiv:1512.02167*. [Online]. Available: http://arxiv.org/abs/1512.02167
[80] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

MARYAM RAHNEMOONFAR (Member, IEEE) received the Ph.D. degree in computer science from the University of Salford, Manchester, U.K. She is currently an Associate Professor and the Director with the Computer Vision and Remote Sensing Laboratory (Bina Lab), UMBC. Her research interests include deep learning, computer vision, data science, AI for social good, remote sensing, and document image analysis, specifically focuses on developing novel machine learning and computer vision algorithms for heterogenous sensors, such as radar, sonar, multi-spectral, and optical. Her research has been funded by several awards, including the NSF Bigdata Award, the Amazon Academic Research Award, the Amazon Machine Learning Award, Microsoft, and IBM.

TASHNIM CHOWDHURY (Graduate Student 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, OH, USA, in 2016. He is currently pursuing the Ph.D. degree in information systems with the University of Maryland, Baltimore County, USA. His research interests include deep learning, machine learning, semantic segmentation, few shot learning, meta learning, and bayesian learning.

ARGHO SARKAR received the B.S. degree in applied statistics from the University of Dhaka, Bangladesh. He is currently pursuing the Ph.D. degree in information systems with the University of Maryland, Baltimore County, USA. His research interest includes developing algorithms for multimodal application, such as visual question answering and image captioning for medical and climate issues.

DEBVRAT VARSHNEY received the B.E. degree in electronics and instrumentation from the Birla Institute of Technology and Science (BITS) Pilani, India, and the M.Sc. degree in Earth observation from a joint collaboration between the University of Twente, The Netherlands, and Indian Space Research Organisation (ISRO). He is currently pursuing the Ph.D. degree in artificial intelligence (AI) with the University of Maryland, Baltimore County, USA, where he is building deep learning algorithms for remotely sensed images. He is currently an experienced Software Engineer. His research interests include semi-supervised learning and physics guided neural networks to efficiently process large climate datasets.

MASOUD YARI received the Ph.D. degree in applied mathematics from Indiana University, Bloomington, IN, USA, in 2008. He is currently a Research Professor with the Department of Information Systems, University of Maryland, Baltimore County, MD, USA. His research interests include mathematical foundations of machine learning, computer vision, partial differential equations, and dynamical systems.

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, in 1980, 1989, and 1992, respectively, where she was a Rockwell International Doctoral Fellow. She is currently a Raytheon Professor of computer science and engineering with Texas A&M University. She is a founder of the fields of rescue robots and human–robot interaction. She has over 100 publications including the best selling textbook, *Introduction to AI Robotics* (MIT Press, 2000). Her research interests include artificial intelligence, human–robot interaction, and heterogeneous teams of robots.