

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

Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debraj Varshney, Masoud Yari
 University of Maryland Baltimore County
 Baltimore, Maryland

`maryam@umbc.edu, tchowdhury@umbc.edu, asarkar2@umbc.edu, dvarshney@umbc.edu, yari@umbc.edu`

Robin Murphy
 Texas A&M University
 College Station, Texas
`murphy@cse.tamu.edu`

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 have 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.

1. Introduction

Understanding of a visual scene from images has the potential to advance many decision support systems. The purpose of scene understanding is to classify the overall category of scene as well as constituting interrelationship among different object classes at both instance and pixel level. Recently, several datasets [19, 48, 23] have been presented to study different aspects of scenes by implementing many computer vision tasks. A major factor in success of most of the deep learning algorithms is the availability of large-scale dataset. Publicly available ground imagery datasets such as ImageNet[19], Microsoft COCO[48], PASCAL VOC[23], Cityscapes[15] accelerate the advanced development of current deep neural networks, but the annotation of aerial imagery is scarce and more tedious to obtain.

Aerial scene understanding dataset are helpful for urban management, city planning, infrastructure maintenance, damage assessment after natural disasters, and high definition (HD) maps for self-driving cars. Existing aerial datasets, however, are limited mainly to classification [29, 44] or semantic segmentation [29, 60] of few individual classes such as roads or buildings. Moreover, all of these datasets are collected in normal conditions and computer vision algorithms are mainly developed for normal looking objects. Most of these datasets do not address the unique challenges in understanding post disaster scenarios as a task for disaster damage assessment. For quick response and recovery in large scale after a natural disaster such as hurricane, wildfire, and extreme flooding access to 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.



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

Although several datasets [8, 7, 30, 66] are provided for post disaster damage assessments, they have numerous issues to tackle. Most of those datasets contain satellite images and images collected from social media. Satellite images are low in resolution and captured from high altitude. They are affected from several noises including clouds and smokes. Moreover, deploying satellites and collecting images from these are costly. On the other hand, images posted on social media are noisy and not scalable for deep learning models. To address this issues, our dataset, *FloodNet*, provides high resolution images taken from low altitude. These characteristics of *FloodNet* brings more clarity to scenes and thus help deep learning models in making more accurate decisions regarding post disaster damage assessment. In addition, most of tasks considering natural disaster datasets are restricted to mainly classification and object detection. Our dataset offers advanced computer vision challenges namely semantic segmentation and visual question answering besides classification. All these three computer vision tasks can provide assistance in complete understanding of a scene and help rescue team to manage their operation efficiently during emergencies. Figure 1 shows sample annotations offered by FloodNet.

Our contribution is two folds. First we introduce a high resolution UAV imagery named *FloodNet* for post disaster damage assessment. Secondly, we compare the performance of sevral classification, semantic segmentation and visual question answering on our dataset. To the best of our knowledge, this is the first VQA work focused on UAV

imagery for any disaster damage assessment.

The reminder of this paper is organized as follows: it begins with highlighting the existing datasets for natural disaster, semantic segmentation, and visual question answering in section 2. Next, section 3 describes the *FloodNet* dataset including its collection and annotation process. Section 4 describes the experimental setups for all three aforementioned tasks along with complete result analysis of corresponding tasks. Finally section 5 summarizes the results including conclusion and future works.

2. 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.

2.1. Datasets

Natural disaster dataset can be initially classified into two classes: A) Non-imaging dataset (text, tweets, social media post) [35, 58] and B) Imaging datasets [60, 29, 12]. Based on the image capture position existing imaging natural disaster datasets can be further classified into three classes: B1) Ground-level images [54], B2) Satellite imagery [12, 29, 22, 17, 14, 60], and B3) Aerial imagery [44, 78, 25]. Recently several datasets have been introduced by researchers for natural disaster damage assessment. Nguyen et al. proposed an extension of AIDR system [53] to collect data from social media in [54]. AIST Build-

Table 1. A brief summary of existing datasets.

Dataset	Types of Images	UAV imagery	Post Disaster	Resolution of Images	Classification	Semantic Segmentation	VQA
ImageNet [19]	Real-world images	No	No	average 400 × 350	✓	✗	✗
Cityscapes [48]	Real-world images	No	No	1280 × 720	✗	✓	✗
DAQUAR [51]	Real-world images	No	No	640 × 480	✗	✗	✓
COCO-QA [57]	Real-world images	No	No	640 × 480	✗	✗	✓
COCO-VQA [4]	Real world images, abstract cartoon images	No	No	640 × 480	✗	✗	✓
Visual Genome [40]	Real-world images	No	No	varies in size	✗	✗	✓
Visual7W [79]	Real-world images	No	No	varies in size	✗	✗	✓
TDIUC [37]	Real-world images	No	No	varies in size	✗	✗	✓
CLEVR [36]	Geometrical Shape	No	No	320 × 240 (in default settings)	✗	✗	✓
PATHVQA [33]	Medical Images	No	No	Varies in size	✗	✗	✓
VQA-MED [2]	Medical Images	No	No	Varies in size	✗	✗	✓
Nguyen et al. [53]	Post Disaster Images	No	Yes	Varies in size	✓	✗	✗
ABCD [25]	Pre and Post Disaster Images	No	Yes	Varies in size	✓	✗	✗
SpaceNet + Deepglobe [22]	Pre and Post Disaster Images	No	Yes	Varies in size	✗	✓	✗
Chen et al. [12]	Post Disaster Images	No	Yes	Varies in size	✗	✗	✗
OSCD [17]	Urban Change Images	No	No	Varies in size	✗	✗	✗
fMoW [14]	Pre and Post Disaster Images	No	Yes	Varies in size	✓	✗	✗
AIDER [44]	Post Disaster Images	Yes	Yes	Varies in size	✓	✗	✗
Rudner et al. [60]	Post Disaster Images	No	Yes	Varies in size	✗	✓	✗
xBD [29]	Pre and Post Disaster Images	No	Yes	Varies in size	✓	✓	✗
ISBDA [78]	Post Disaster Images	Yes	Yes	Varies in size	✗	✗	✗
FloodNet (Ours)	Post Disaster Images	Yes	Yes	4000×3000	✓	✓	✓

ing Change Detection (ABCD) dataset has been proposed in [25] which includes aerial post tsunami images to identify whether the buildings have been washed away. A combination of SpaceNet [16] and DeepGlobe [18] was presented in [22] and a segementation model was proposed to detect changes in man-made structures to estimate the impact of natural disasters. Chen et al. in [12] 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 [17] 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 [14]. Aerial Image Database for Emergency Response (AIDER) is proposed by Kyrou et al. in [44] for classification of UAV imagey. Rudner et al. [60] propose a satellite imagery collected from Sentinel-1 and Sentinel-2 satellites for semantic segmentation of flooded buildings. Gupta et al. proposed xBD [29] 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 [78] 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 resulting 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 are primary satellite imageries. 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 [54]. These imageries lack geo location tags [78] and suffers from data scarcity for deep learning training [66]. Although some aerial datasets [44, 78] are prepared using UAVs, these datasets lack low altitude high resolution images. AIDER [44] 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 [78] 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 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.

2.2. Algorithms

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

2.2.1 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 AlexNet [43] which achieved state-of-the-art performance on the ImageNet [20] 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[42, 41], MNIST[47], and FashionMNIST [67].

This led to a rise in networks such as [63], [32], [64], [13], [34] etc., where the network architectures were exper-

mented 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 aerial or space borne sensors.

In this regard, some image classification networks have been explored for the purpose of post-disaster damage detection [45, 53, 5, 61, 76]. [53] used crowd sourced images from social media which captured disaster sites from the ground level. [5] used a Support Vector Machine on top of a Convolutional Neural Network (CNN) followed by a Hidden Markov Model post-processing to detect avalanches. [61] compared [63] and [32] for fire detection, but then again the dataset used contained images taken by hand-held cameras on the ground. [76] developed a novel algorithm which focused on wildfire detection through UAV images. [45] have done extensive work by developing a CNN for emergency response towards fire, flood, collapsed buildings, and crashed cars. Our paper can contribute in this domain by providing multi feature flooded scenes that can inspire the efficient training of more neural networks.

2.2.2 Semantic segmentation

Semantic segmentation is one of the prime research area in computer vision and an essential part of scene understanding. Fully Convolutional Network (FCN) [49] 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 [74] or DeepLab [9, 10] perform spatial pyramid pooling [28, 46] 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 different scale global context. Some notable works using this architecture are [10, 59]. On the other hand, there are models [75, 73, 24] 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 [22] perform previously proposed semantic segmentation [21] 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 [56] 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 [60] for rapid

segmentation of flooded buildings. Gupta et al. propose a DeepLabv3 [10] and DeepLabv3+ [11] inspired RescueNet in [31] 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 [15, 52]. 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 [55], one pyramid pooling module based network PSPNet [74], and the last network model DeepLabv3+ [11] employs both encoder-decoder and pyramid pooling based module.

2.2.3 Visual Question Answering

Many researchers proposed several datasets and methods for Visual Question Answering task. However, there are no such datasets apt for training and evaluating VQA algorithms regarding disaster damage assessment tasks.

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 [4, 69, 38, 27] using deep learning-based multimodal methods [50, 68, 26, 3, 70, 72, 6, 39, 71]. 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[62], COCO-VQA [4], Visual Genome[40], Visual7W [79]. In addition Path-VQA [33] and VQA-MED [2] are medical images for which VQA algorithms are also considered. 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[69], and MFB with Co-Attention[71] network on our dataset.

3. 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 con-

tains 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. Such images are very high in resolution, making them unique compared to other data sets 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 visual question answering, these attributes are considered.

3.1. 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 3200 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 take 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 [1] for classification and semantic segmentation. 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	275	3573
Building-non-flooded	1272	5373
Road-flooded	335	649
Road-non-flooded	1725	3135
Vehicle	1105	6058
Pool	676	1421
Tree	2507	25889
Water	1262	1784

3.2. 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 object 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, $\sim 11,000$ 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.

3.2.1 Types of Question

Questions are divided into a three-way question group, namely “*Simple Counting*”, “*Complex Counting*”, 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 belongs 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

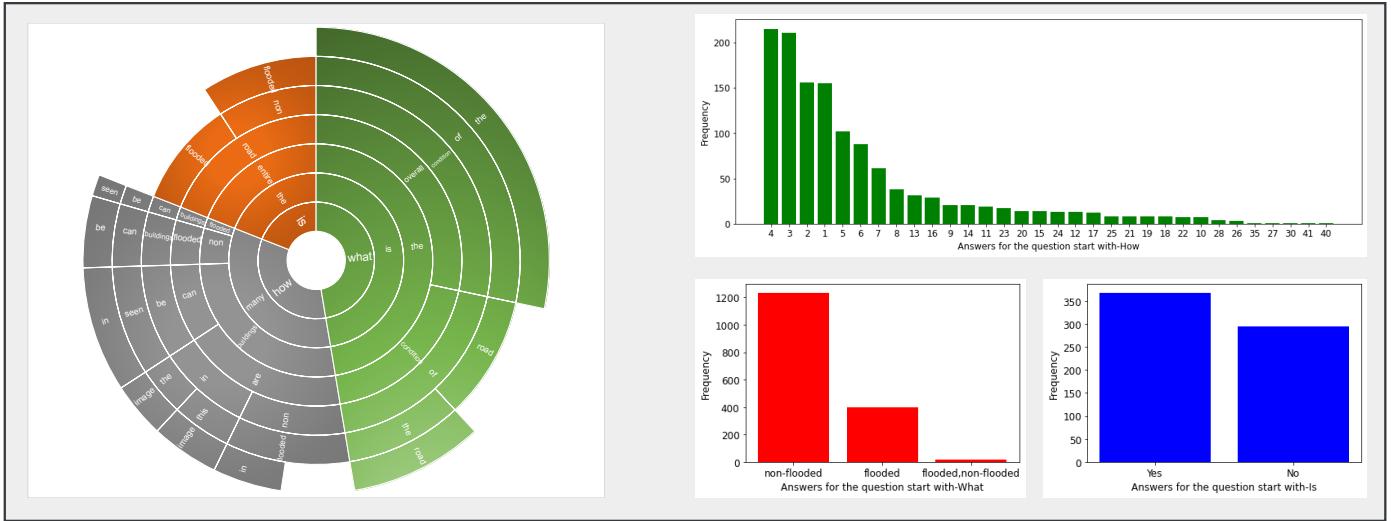


Figure 2. VQA Data Statistics: Left figure represents the distribution of the question pattern based on starting word; Right-top, Right-bottom figures describes the distribution of possible answers for different question types

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 the road. This type of question is divided into three sub-categories. One category deals with the condition of road by asking questions such as “What is the condition of the road?”. Second one seeks the condition of the entire image by asking questions like “What is the overall condition of the entire image?”. “Yes/No” type question is categorised as the third sub-category of the *Condition Recognition*. “Is the road flooded?”, “Is the road non-flooded” are some of the examples from this sub-category. Starting word for this type of question is either “Is” or “What”.

3.2.2 Types of Answer

Table 3. Possible Answers for Three Types of Questions

Question Type	Possible Answer
Simple Counting	{1,2,3,4...}
Complex Counting	{1,2,3,4...}
Condition of Road	Flooded , Non-Flooded, Flooded & Non-Flooded
(sub-category of Condition Recognition)	Condition of Entire Image
(sub-category of Condition Recognition)	Flooded , Non-Flooded
Yes/No-Type Question	Yes, No
(sub-category of Condition Recognition)	

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, 30, 41, 40’ are the less frequent answers. For *Condition Recognition* problem, ‘non-flooded, yes’ are the most common answers.

4. Experiments

To understand the usability of these images for flood detection, we majorly carry out three tasks, which are Image Classification, Semantic Segmentation, and Visual Question Answering (VQA). We begin with classifying the FloodNet data into Flooded and Non-Flooded images, then we detect specific regions of flooded buildings, flooded roads, vehicles etc. through semantic segmentation networks. Finally, we carry out VQA on this dataset. For all of our tasks, we use NVIDIA GeForce RTX 2080 Ti GPU with an Intel Core i9 processor.

For image classification, we used three state-of-the-art networks i.e. InceptionNetv3 [65], ResNet50 [32], and Xception [13] as base models to classify the images into Flooded and Non-Flooded categories. 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 Inception-Net 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 Relu Activation, and finally by two neurons with Softmax activation. We

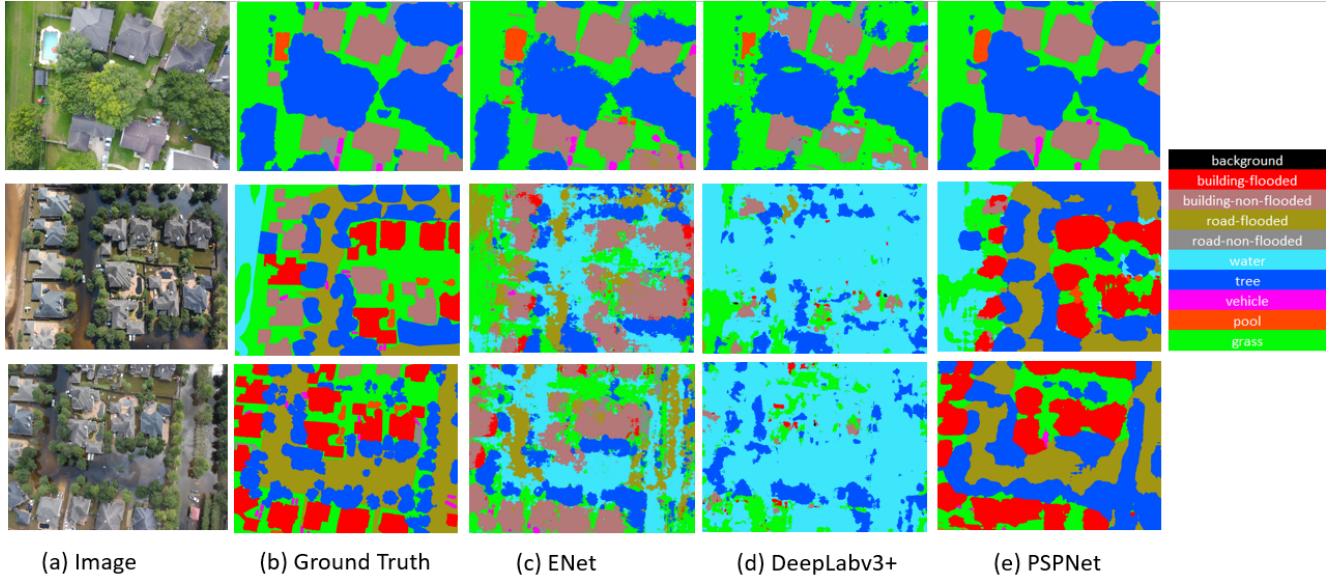


Figure 3. Visual comparison on FloodNet test set for Semantic Segmentation.

Table 4. Per-class results on FloodNet testing set.

Method	Building Flooded	Building Non Flooded	Road Flooded	Road Non Flooded	Water	Tree	Vehicle	Pool	Grass	mIoU
ENet[55]	6.94	47.35	12.49	48.43	48.95	68.36	32.26	42.49	76.23	42.61
DeepLabV3+[11]	32.7	72.8	52.00	70.2	75.2	77.00	42.5	47.1	84.3	61.53
PSPNet[74]	68.93	89.75	82.16	91.18	92.00	89.55	46.15	64.19	93.29	79.69

initialized our networks with ImageNet [20] weights and trained them for 30 epochs, with 20 steps for every epoch, using binary cross entropy loss.

For semantic segmentation, we implemented three methods, i.e. PSPNet [74], ENet [55], and DeepLabv3+ [11]; 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 IoU (mIoU).

For Visual Question Answering, simple baselines (concatenation/element-wise product of image and text features) and Multimodal Factorized Bilinear (MFB) with co-

attention [71], Stacked Attention Network [69] have been considered for this study. All of these models are configured according to our dataset. For image and question feature extraction, respectively, VGGNet (VGG 16) and Two-Layer LSTM are taken into account. Feature vector from last pooling layer of the VGGNet and 1024-D vector from the last word of Two-Layer LSTM are considered as the image and question vectors respectively. Dataset is splitted into training, validation and testing data. All the images are resized to 224×224 and questions are tokenized. 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.

4.1. Image Classification Analysis

The classification accuracies of the three networks are shown in Table 6. From this table, it can be seen that the highest performance on the test set was given by ResNet. The residual architecture of ResNet has successfully helped in classifying the test images into Flooded and Non-Flooded, as compared to the other networks. Even though Xception and InceptionNet have a much wider architecture and show higher classification accuracy on ImageNet data, this is not the case for FloodNet dataset.

Table 5. Accuracy table for Baseline VQA Algorithms

Method	Data Type	Overall Accuracy	Counting Problem		Condition Recognition		
			Accuracy for 'Simple Counting'	Accuracy for 'Complex Counting'	Accuracy for 'Yes/No'	Accuracy for "Entire Image Condition"	Accuracy for "Road Condition"
Concatenation of Features [77]	Validation	0.41	0.04	0.03	0.017	0.86	0.9
	Testing	0.42	0.04	0.03	0.17	0.86	0.9
Element-wise Multiplication of Features [44]	Validation	0.69	0.28	0.27	0.86	0.96	0.97
	Testing	0.68	0.25	0.21	0.84	0.96	0.97
SAN [69]	Validation	0.63	0.34	0.28	0.51	0.95	0.97
	Testing	0.63	0.26	0.24	0.54	0.94	0.97
MFB with Co-Attention [71]	Validation	0.72	0.31	0.28	0.98	0.96	0.97
	Testing	0.73	0.29	0.26	0.99	0.97	0.99

Therefore, networks which give high accuracy on everyday images such as those of ImageNet can not 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 6. Classification accuracy of three state-of-the-art networks on FloodNet dataset

Model	Training Accuracy	Test Accuracy
InceptionNetv3[65]	99.03 %	84.38%
ResNet50[32]	97.37%	93.69%
Xception[13]	99.84 %	90.62%

4.2. 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. Then 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, still their performance on detecting flooded building and flooded roads are still low, since distinguishing between flooded and non-flooded objects heavily depend on respective contexts of the classes.

4.3. Visual Question Answering Performance Analysis

From the Table 5, we can identify that counting problem (simple and complex) is very challenging compare to task of condition recognition. Many objects are very small which makes it very difficult even for human to count. Accuracy for '*Condition Recognition*' category is high. 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. MFB with co-attention [71] outperforms all the other methods for all types of question.

5. 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 visual question answering 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.

6. Acknowledgment

This work is partially supported by Microsoft and Amazon.

References

- [1] V7 darwin. <https://www.v7labs.com/darwin>. Accessed: 2020-11-11. 5
- [2] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *CLEF2019 Working Notes*, 2019. 3, 4
- [3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 4
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Visual question answering. In *ICCV*, 2015. 3, 4, 8
- [5] Mesay Bejiga, Abdallah Zeggada, Abdelhamid Nouffidj, and Farid Melgani. A convolutional neural network approach for assisting avalanche search and rescue operations with uav imagery. *Remote Sensing*, 9(2):100, Jan 2017. 4
- [6] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017. 4
- [7] Bischke Benjamin, Helber Patrick, Zhao Zhengyu, Borth Damian, et al. The multimedia satellite task at mediaeval 2018: Emergency response for flooding events. 2018. 2
- [8] Benjamin Bischke, Patrick Helber, Christian Schulze, Venkat Srinivasan, Andreas Dengel, and Damian Borth. The multimedia satellite task at mediaeval 2017. In *MediaEval*, 2017. 2
- [9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 4
- [10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 4
- [11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 4, 7
- [12] Sean Andrew Chen, Andrew Escay, Christopher Haberland, Tessa Schneider, Valentina Staneva, and Youngjun Choe. Benchmark dataset for automatic damaged building detection from post-hurricane remotely sensed imagery. *arXiv preprint arXiv:1812.05581*, 2018. 2, 3
- [13] François Fleuret. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 3, 6, 8
- [14] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 2, 3
- [15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 4
- [16] DigitalGlobe CosmoQWorks. Nvidia: Spacenet on amazon web services (aws) datasets: The spacenet catalog. 3
- [17] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2115–2118. IEEE, 2018. 2, 3
- [18] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raska. Deepglobe 2018: A challenge to parse the earth through satellite images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 172–1729. IEEE, 2018. 3
- [19] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1, 3
- [20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 7
- [21] Jigar Doshi. Residual inception skip network for binary segmentation. In *CVPR Workshops*, pages 216–219, 2018. 4
- [22] Jigar Doshi, Saikat Basu, and Guan Pang. From satellite imagery to disaster insights. *arXiv preprint arXiv:1812.07033*, 2018. 2, 3, 4
- [23] 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. *International Journal of Computer Vision*, 111(1):98–136, Jan. 2015. 1
- [24] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 4
- [25] Aito Fujita, Ken Sakurada, Tomoyuki Imaizumi, Riho Ito, Shuhei Hikosaka, and Ryosuke Nakamura. Damage detection from aerial images via convolutional neural networks. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pages 5–8. IEEE, 2017. 2, 3
- [26] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016. 4
- [27] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6639–6648, 2019. 4

- [28] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1458–1465. IEEE, 2005. 4
- [29] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 10–17, 2019. 1, 2, 3
- [30] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019. 2
- [31] Rohit Gupta and Mubarak Shah. Rescuenet: Joint building segmentation and damage assessment from satellite imagery. *arXiv preprint arXiv:2004.07312*, 2020. 4
- [32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 4, 6, 8
- [33] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 3, 4
- [34] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3
- [35] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):1–38, 2015. 2
- [36] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 3
- [37] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973, 2017. 3
- [38] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018. 4
- [39] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. 4
- [40] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 3, 4
- [41] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3
- [42] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 3
- [43] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 3
- [44] Christos Kyrkou and Theocharis Theοcharides. Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles. In *CVPR Workshops*, pages 517–525, 2019. 1, 2, 3
- [45] Christos Kyrkou and Theocharis Theοcharides. Emergencynet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:1687–1699, 2020. 4
- [46] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2169–2178. IEEE, 2006. 4
- [47] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010. 3
- [48] Tsung-Yi Lin, M Maine, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C Lawrence Zitnick, and Piotr Dollar. Microsoft coco: Common objects in context. arxiv, 2015. 1, 3
- [49] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 4
- [50] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, 2016. 4
- [51] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pages 1682–1690, 2014. 3
- [52] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014. 4
- [53] Dat Tien Nguyen, Firoj Alam, Ferda Ofli, and Muhammad Imran. Automatic image filtering on social networks using deep learning and perceptual hashing during crises. *arXiv preprint arXiv:1704.02602*, 2017. 2, 3, 4
- [54] Dat T Nguyen, Ferda Ofli, Muhammad Imran, and Prasenjit Mitra. Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 569–576, 2017. 2, 3

- [55] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016. 4, 7
- [56] Maryam Rahneemoonfar, Robin Murphy, Marina Vicens Miquel, Dugan Dobbs, and Ashton Adams. Flooded area detection from uav images based on densely connected recurrent neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1788–1791. IEEE, 2018. 4
- [57] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961, 2015. 3
- [58] Christian Reuter and Marc-André Kaufhold. Fifteen years of social media in emergencies: a retrospective review and future directions for crisis informatics. *Journal of Contingencies and Crisis Management*, 26(1):41–57, 2018. 2
- [59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4
- [60] Tim GJ Rudner, Marc Rußwurm, Jakub Fil, Ramona Pelich, Benjamin Bischke, Veronika Kopačková, and Piotr Biliński. Multi3net: segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 702–709, 2019. 1, 2, 3, 4
- [61] Jivitesh Sharma, Ole-Christoffer Granmo, Morten Goodwin, and Jahn Thomas Fidje. Deep convolutional neural networks for fire detection in images. In *International Conference on Engineering Applications of Neural Networks*, pages 183–193. Springer, 2017. 4
- [62] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 746–760, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 4
- [63] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 4
- [64] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 3
- [65] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 6, 8
- [66] Ethan Weber, Nuria Marzo, Dim P. Papadopoulos, Aritro Biswas, Agata Lapedriza, Ferda Ofli, Muhammad Imran, and Antonio Torralba. Detecting natural disasters, damage, and incidents in the wild. In *The European Conference on Computer Vision (ECCV)*, August 2020. 2, 3
- [67] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashionmnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 3
- [68] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*. Springer, 2016. 4
- [69] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016. 4, 7, 8
- [70] Dongfei Yu, Jianlong Fu, Xinmei Tian, and Tao Mei. Multi-source multi-level attention networks for visual question answering. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2019. 4
- [71] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017. 4, 7, 8
- [72] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959, 2018. 4
- [73] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018. 4
- [74] Hengshuang Zhao, Jianping Shi, Xiaojian Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 4, 7
- [75] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 267–283, 2018. 4
- [76] Yi Zhao, Jiale Ma, Xiaohui Li, and Jie Zhang. Saliency detection and deep learning-based wildfire identification in uav imagery. *Sensors*, 18(3):712, 2018. 4
- [77] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015. 8
- [78] Xiaoyu Zhu, Junwei Liang, and Alexander Hauptmann. Msnet: A multilevel instance segmentation network for natural disaster damage assessment in aerial videos. *arXiv preprint arXiv:2006.16479*, 2020. 2, 3
- [79] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016. 3, 4