


Article

# Feasibility of Visual Question Answering (VQA) for Post-Disaster Damage Detection Using Aerial Footage †

Rafael De Sa Lowande <sup>1</sup> and Hakki Erhan Sevil <sup>2,\*</sup> 

<sup>1</sup> Electrical and Computer Engineering, University of West Florida, Pensacola, FL 32514, USA; rl64@students.uwf.edu

<sup>2</sup> Intelligent Systems & Robotics, University of West Florida, Pensacola, FL 32514, USA

\* Correspondence: hsevil@uwf.edu

† This paper is an extended version of our papers published in Clevenger, A., Lowande, R., Sevil, H.E., Mahyari, A., "Towards UAV-Based Post-Disaster Damage Detection and Localization: Hurricane Sally Case Study," AIAA SciTech 2022, San Diego, CA, USA, 3–7 January 2022, AIAA-2022-0788, and Lowande, R., Mahyari, A., Sevil, H.E., "Post-Disaster Damage Detection using Aerial Footage: Visual Question Answering (VQA) Case Study," IEEE Applied Imagery Pattern Recognition (AIPR 2022), Cosmos Club, Washington, DC, USA, 11–13 October 2022.

**Abstract:** Natural disasters are a major source of significant damage and costly repairs around the world. After a natural disaster occurs, there is usually a significant amount of damage, and with that, there are also a lot of costs involved with repairing and aiding all the people involved. In addition, the occurrence of natural phenomena has increased significantly in the past decade. With that in mind, post-disaster damage detection is usually performed manually by human operators. Taking into consideration all the areas one has to closely look into, as well as the difficult terrain and places with hard access, it becomes easy to understand how incredibly difficult it is for a surveyor to identify and annotate every single possible damage out there. Because of that, it has become essential to find new creative solutions for damage detection and classification in the case of natural disasters, especially hurricanes. This study focuses on the feasibility of using a Visual Question Answering (VQA) method for post-disaster damage detection, using aerial footage taken from an Unmanned Aerial Vehicle (UAV). Two other approaches are also utilized to provide comparison and to evaluate the performance of VQA. Our case study on our custom dataset collected after Hurricane Sally shows successful results using VQA for post-disaster damage detection applications.

**Keywords:** visual question answering; post-disaster; damage detection; aerial footage



**Citation:** Lowande, R.D.S.; Sevil, H.E. Feasibility of Visual Question Answering (VQA) for Post-Disaster Damage Detection Using Aerial Footage. *Appl. Sci.* **2023**, *13*, 5079. <https://doi.org/10.3390/app13085079>

Academic Editor: Andrea Prati

Received: 8 March 2023

Revised: 18 April 2023

Accepted: 18 April 2023

Published: 19 April 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

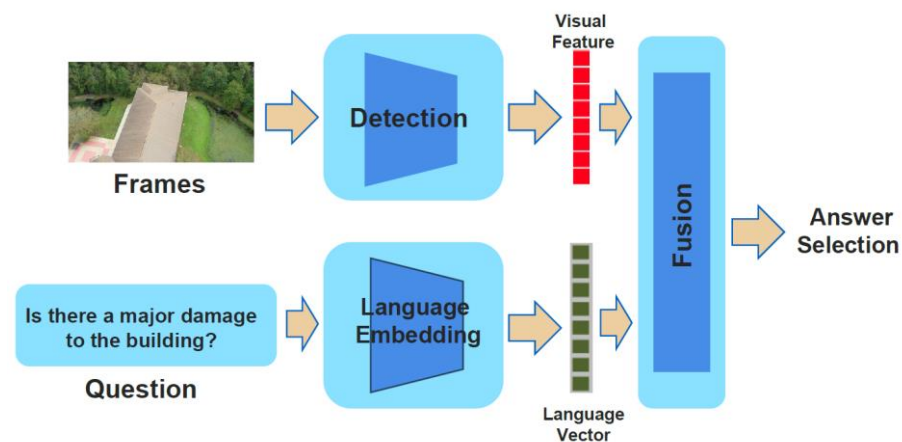
## 1. Introduction

It has been clear in the past few decades how natural disasters are a major source of significant damage and costly repairs around the world. In 2020 alone, more than USD 43 billion of damage resulted from the Atlantic hurricane season in North America [1]. In perspective, according to *The Wall Street Journal*, 31% of all hurricane damages from 1980 to 2018 occurred in 2017, with a total of USD 268 billion in damages [2]. This shows how the occurrence of natural disasters is constantly increasing, especially in the last decade. Taking this information into consideration, the need for the means to quickly detect damages and respond to these disasters has also significantly increased recently.

Post-disaster damage detection is usually performed manually by human surveyors. Considering all the areas one has to closely look into, as well as the difficult terrain and places with hard access, it becomes easy to understand how incredibly difficult it is for a surveyor to identify and annotate every single possible amount of damage out there. It is also reasonably understandable to assume that the surveyor will miss some type of information in the field that could be essential. Hence, this method of damage identification

has become obsolete, making it a slow process and an inefficient and inconsistent way of securing human resources.

In this study, we propose to utilize an Unmanned Aerial Vehicle (UAV) with an attached standard color camera to capture footage of the post-disaster conditions of structures using computer vision techniques. In the literature, UAVs have been used frequently for damage detection and identification applications after natural disasters. In this study, we use a quadrotor UAV, as they possess higher mobility that allows the capturing of different points of view for the scene of interest. The main focus of our study is to perform proof-of-concept analysis for the Visual Question Answering (VQA) approach. The concept of VQA is depicted in Figure 1. For the “Detection” part, we utilize a Convolutional Neural Network (CNN), and for the “Language Embedding” part, we utilize a Bag of Words (BoW).



**Figure 1.** The pipeline of Visual Question Answering (VQA)-based post-disaster damage detection.

In order to evaluate the feasibility and performance of VQAs, we also present damage detection approaches, using a Cascade Classifier (CC) that is described in Section 3, and the CNN (only CNN, without BoW) that is described in Section 4. Our goal is to analyze and provide the positive and negative aspects of each approach by comparing the results from each one. This includes a detailed analysis of each model and provides reasoning for using one model instead of the other when trying to perform the detection in specific scenarios. For that goal, post-disaster damage detection analysis is conducted on University of West Florida (UWF) campus aerial footage that is unique and gathered after Hurricane Sally (Sally-UWF), which was a slow-moving category two hurricane that made a landfall in 2020 around Pensacola, FL. Additionally, custom annotations and images created for roof damages, and three different classes demonstrating low, medium, and heavy damages, as well as the instance and frequency of the presence of roof damages, are introduced. Different approaches are studied throughout this paper in order to have a good comparison base when testing the actual feasibility of implementing the VQA model for damage detection. The original contributions of this study can be summarized as (i) the use of custom aerial footage data taken from the UWF campus after Hurricane Sally, (ii) the development of custom-annotated data for roof damage detection, (iii) the development of the VQA pipeline and implementation on aerial footage data in order to validate its feasibility for post-disaster damage detection, and (iv) performing a comparison between the VQA-, single CNN-, and single CC-based detection methods to highlight the feasibility of the VQA approach. As our aim is to show the proof-of-concept results of our developed pipeline, the comparison of our results with other approaches, such as you only look once (YOLO) models, is not included in the scope of the originally intended contribution in this article, and is left for a future work.

The remainder of the paper is organized as follows. Related work on the CNN, CC, and VQA is presented in the next section. The first approach, CC, is presented and analyzed

in Section 3. The CNN approach is presented in Section 4. In Section 5, the Visual Question Answering is presented. Section 6 presents the analyses and comparison between all the approaches of this research study, their pros and cons, and provides the justification for their use for post-disaster damage detection. In Section 7, the conclusions of this study are presented and the planned improvements are listed.

## 2. Related Work

In the literature, image processing, computer vision, and pattern recognition-related research studies have received a lot of attention in recent years due to advancements in algorithms and equipment used in their application, which varies from object detection [3] and object tracking [4], to 3D modeling [5]. The Cascade Classifier, the first method we analyzed for object detection, has shown significant success for object detection tasks in the past. Viola and Jones [6] present a comparison of several different cascade of classifier detectors with high detection rates for face detection, with low false-positive rates, which is typically a significant issue with cascade of classifier models. A similar study performed by Lienhart et al. [7] also presents successful results for different classifier boosting strategies applied to cascade of classifier type models trained for face detection. Wang et al. [8] expand the usage of Cascade Classifiers to the general case with the PASCAL VOC datasets (20 object classes) and the ImageNet dataset (200 object classes).

CNNs have become widely popular for carrying out object detection tasks in recent years. A notable study presented by Zhu et al. [9] for roof damage detection uses their own CNN model to perform the detection with excellent accuracy. A more general building damage study by Nex et al. [10] utilizes CNNs with a morphological filter method for damage candidate region proposals. Pi et al. [11] compare a number of CNN architectures for post-hurricane damage detection with a high mean average precision. Some recent methods have combined the cascading strategy with CNNs, as presented by Cai and Vasconcelos [12]. In their paper, the researchers use a cascading region with a sequentially higher intersection over union thresholds to filter out false-positive samples.

Looking into the Visual Question Answering (VQA) approach, there has been considerable research conducted on this new model recently. Studies [13–21] have made significant efforts in order to develop and analyze the approach. These studies propose different approaches for the union of semantic images and question features.

VQA is often used for extracting information from remotely sensed data [22]. Lobry et al. introduced a model based on deep learning to analyze images and corresponding questions. They used both recurrent and convolutional neural networks in their VQA model, and they had promising results to answer specific questions regarding buildings in the images [22]. Silva et al., on the other hand, presented a human recognition application in VQA [23]. In their study, they introduced a novel feature extractor in order to estimate 2D poses of humans, which then leads to an estimation of action by humans. Their results showed a 62.03% mean average precision for a very large video dataset used in their study [23]. In a similar study, Vo et al. introduced VQASTO, which is a task ontology-based Visual Question Answering method for action surveillance [24]. In that study, two subtasks are used: skeleton-based action recognition and pose estimation/tracking. With these subtasks, the system maps question sentences to the corresponding action and outputs an answer [24].

One significant application area of VQA is the medical field [25]. The idea for models in this field is to answer medical questions based on the visual content of images, for instance radiology images [25]. Abacha et al. presented the results of 17 teams that tried to achieve the best VQA results, and the team with best result had an accuracy of 62.4% on 4200 radiology images. VQA also is used for specific application purposes in the literature, such as the detection of occupancy spaces for parking. Hela et al. introduced a VQA model, which is built by manually labeling questions and answers on the CNRPark + Ext dataset [26]. In their study, they ended up with an average accuracy of 94.31% for detecting parking spots [26].

Although all these studies are presented in the literature, there are not many studies that address the use of VQA paired with UAVs in order to identify damages caused by natural disasters, and it is still an open area of research.

### 3. A Cascade Classifier Model Using Haar Features

The first approach that is analyzed in this paper is the Cascade Classifier (CC). Proposed by Paul Viola and Michael Jones in 2001 [6], the Cascade Classifier is a machine-learning method that uses positive and negative images to train its model and obtain the final detection. Positive images are considered to be all the images containing the object the model is searching for, while negative images are considered to be all the images that do not contain the object that the model is searching for.

The algorithm for the Cascade Classifier requires a significant number of positive images and negative images, and separates them into two files. It then extracts the features from each file; for this paper, the Haar features are used in order to detect and extract the features from the images. The Haar features are a sequence of rescaled square-shape functions first introduced in the literature by Alfred Haar in 1909 [6]. The next step in this approach is to define a threshold. In a perfect scenario, when detecting an object of interest in an image, the algorithm would return a perfect match. However, since it is very unlikely that the model will find the final result with 100% certainty every time the object is in the image, a threshold is determined. For example, setting the threshold to 0.6 (60%) would mean that every time the algorithm returns a value equal to or greater than 0.6, the model would consider that the object has been detected. On the other hand, if it returns a value below 0.6, that means the model has not detected the object it has been looking for. Setting the threshold too low could lead to many false positives, while setting the threshold too high could lead to false negatives.

After identifying all the features, the algorithm starts a process that applies the features to all the collected images, then the algorithm then classifies each image into positive or negative, with the positive ones being the ones that contain the “object” and negative ones being the ones that do not. The more data used, the better the prediction will be, and therefore, the final results will be more accurate.

In our study, it is expected that the majority of areas in an image will not have visible damages. Therefore, the Cascade Classifier is modified to speed up the process of detection, as well as to locate the detected damage inside a positive image. Thus, instead of applying all the features to an image, the features are grouped into different stages of the classifier and applied individually. If an image fails the first “layer” of features, that image is discarded. If an image passes the first stage, it means that it is possible that the image being analyzed potentially has the damage the algorithm is trying to find. Therefore, the second stage is applied to it, followed by the third stage, and so on, until the algorithm can detect, with certainty, if there is damage on that image or not. For the training of our model, it is necessary to first separate the training data into positive and negative images.

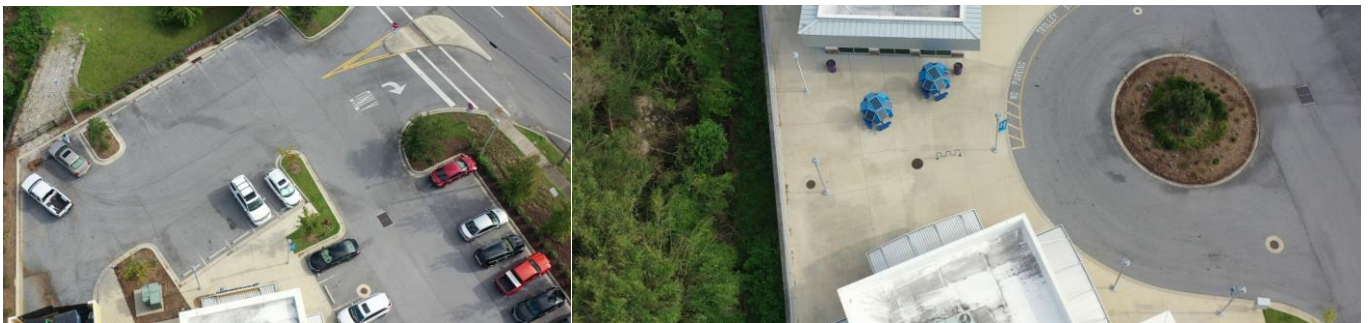
The Sally-UWF dataset consists of aerial footage that was recorded following the hurricane using an UAV with an attached camera. The videos were captured at  $3840 \times 2160$  p with 24 frames per second. All frames have been downsized to  $1920 \times 1080$  p for analysis in this study. In total, 24 videos ranging from 15 s to 5 min were captured. From this collection, two videos with high damage instance counts were selected and used for testing. Some examples of the positive images obtained from the Sally-UWF dataset are depicted in Figure 2.

Contrary to the positive images, the negative images are all the images that do not contain any type of damage. The same Sally-UWF dataset was used to separate and identify the negative images. Some examples of the negative images obtained from the Sally-UWF dataset are depicted in Figure 3.





**Figure 2.** Examples of the positive damage images for the CC approach.



**Figure 3.** Examples of the negative damage images for the CC approach.

In total, 1000 negative images and 1100 positive images were used. The data were trained in a total of twenty stages, which was the maximum number of training stages possible without leading the algorithm to over-fitting. In the case of object detection, the model prediction consists of two parts: the bounding boxes and the corresponding class label. The bounding box is the area of interest for our model, with a range from  $x_1$  to  $x_2$  and  $y_1$  to  $y_2$  in pixel coordinates. Ideally, this box perfectly surrounds the damage to be detected. Along with each box comes the class label, which is the description of what kind of damage the bounding box represents, as well as the confidence ratings. Another technique was also used in this part to facilitate the identification of damage, which is called grouping rectangles. By grouping all the bounding boxes that are relatively close to each other, it becomes easier and simpler to observe the damage being identified in an image.

For the testing, two different videos including footage of two different buildings at UWF, the Martin Hall and Argo Hall, were studied. The reason behind choosing this specific footage is that both of the buildings had a high number of damage instances. For the analysis, the average detection rate and the average number of false positives detected are calculated. The frames were sampled using evenly spaced time intervals. A prediction is considered a true positive if the bounded area contains at least 50% of the ground truth area for that damage. If the bounded area covers less than 50% of the ground truth area or there is no relevant damage at the bounded location, the prediction becomes a false positive. When a prediction is not made for relevant damage instances, this is a false negative. The analysis for this model as well as the comparison of the results with results from other models is discussed in Section 6.

#### 4. The Convolutional Neural Network Model

The next approach analyzed in this paper is the Convolutional Neural Network (CNN). The goal of the CNN model is to reduce the images into a form in which it is easier to process without losing the features that are critical for providing a good prediction [27]. The basic process of using a neural network starts with building a model, training the network, and finally, testing the network on a validation or on real-world data. The training process takes in some labeled training data and gives a prediction for that data. When the model prediction does not match the expected output, the internal model weights are

adjusted through the process of backpropagation, which is a backwards traversal through the network, updating each layer of weights along the way, with the goal of changing values to provide a better prediction in the next iteration. When enough high-quality, representative training data are used, the network is likely to provide high-quality outputs that match what is desired.

Convolutional neural networks use all of these principles of a neural network, and apply them to image frames. The process of convolution involves using kernels, which are small square matrices with specific values. These kernels are multiplied by the pixel values of every area of an image sequentially to create an output matrix called a “feature map.” Typically, different kernels are used to create different feature maps for each image frame. These feature maps undergo additional processing, such as padding and pooling, which aims to reduce the size of the representation while losing as little accuracy as possible for the further iterations of convolution to strengthen the feature representation. In general, the CNNs do not need to be limited to only one Convolutional Layer. Usually, the first CNN layer is responsible for capturing the Low-Level features such as the colors, gradient orientation, and edges, among others. By adding more layers, the general architecture adapts to High-Level features as well, providing a full network [28]. When the convolutional process is complete, the resulting feature maps are flattened and inputted as vectors into what is essentially a standard neural network for classification purposes, where the actual prediction is made. Furthermore, two extensions of CNN, namely region-based CNN (R-CNN) [29] and Fast R-CNN [30], are presented in the literature.

In this study, the TensorFlow Object Detection API for Python is utilized for damage detection. This API features a suite of convolutional neural network models designed for object detection. All models in the TensorFlow model zoo come pretrained on the COCO 2017 dataset as a starting point, but can simply be re-trained with the appropriate data to fit any object detection task. The Faster R-CNN Inception ResNet V2  $640 \times 640$  p model is used and re-trained on the ISBDA dataset [31]. This model is one of the more accurate in precision in the model zoo, scoring a 37.7% mean average precision score. This accuracy comes at the cost of a fairly long processing time of 206 milliseconds per frame. If this system requires immediate analysis of a UAV’s footage, it could lead to a delay, but in the scenario of a post-processing case, it is not critical. Additional Python libraries used for this study include OpenCV for video reading and writing, as well as NumPy for the image format manipulation.

The developed algorithm consists of three parts: the bounding boxes, the corresponding class label, and the corresponding confidence ratings, similar to the CC approach. For this approach, the Instance Segmentation in Building Damage Assessment (ISBDA) dataset is used [31]. This dataset consists of 1030 total images, 908 of them were selected to be used in our study. The segmentation annotations, damage-bounding box annotations, and house-bounding box annotations are provided with the dataset, but those annotations were not used in our analysis. Instead, these images were re-annotated to surround each damage area. Instances were labeled into three distinct classes: “Light,” “Medium,” and “Heavy,” corresponding to the level of damage present. “Light” damage may refer to single or patch single damage, small debris, water damage, major discoloration, or slight bending of metal roofs. “Medium” damage may refer to exposed wooden portions or open areas of the structure without collapse, major debris, or significant bending of metal roofs. “Heavy” damage may refer to the structural collapse of the roof, the complete removal of the roof, or the destruction of the building as a whole. The trained algorithm was then tested with the Sally-UWF dataset. The complete analysis and results, as well as the comparison with other approaches, are discussed in Section 6.

## 5. The Visual Question Answering Model

The next approach to be analyzed in this study is the Visual Question Answering (VQA). VQA is a relatively more recent method in computer vision literature, and different from previously presented approaches, the VQA method needs to demonstrate a more

profound knowledge of images. It needs to be able to answer completely different questions based on an image. With that, the user does not have to predetermine anymore what the algorithm will be looking for in an image. Using this method, the user can just ask questions based on an image, and the algorithm runs an analysis based on the user's question and builds an answer upon that question, for instance, to find the object or person asked by the question.

A VQA approach integrated with a UAV can be fundamental in the advancement of damage detection and assessment in the case of natural disasters, such as hurricanes. Since aiding damaged areas is an activity that is heavily dependent on real-time evaluation and estimation, the introduction of a VQA model can prove to be essential when dealing with high-risk situations. The VQA model is considered a complicated, multimodal research problem in which the aim is to address an image-specified question [32–34]. VQA can be considered a type of comprehensible approach that differentiates itself from other types of approaches, of which most are based on the identification of images. Since the VQA model needs to have a higher understanding of the attributes of an image, and it also needs to find all the relevant objects based on natural language questions, thus, it can prove to be an important aspect in the support for damage detection and identification after natural disasters. An example of the VQA approach is depicted in Figure 4.



**Figure 4.** Visual Question Answering basic application example.

Considering what is essential after a hurricane passes by, some of the first questions that come to mind are “Is there anyone in the area?” or “How many houses were destroyed?” among other questions. Being able to answer these questions in real-time is one of the many benefits provided by the VQA approach. The same dataset (Sally-UWF) is used in the analysis of VQA. Overall, 1000 frames were selected to be analyzed. In total, 1000 training annotations and 30 training questions were used to compose this approach. Furthermore, two videos with high damage instance counts were selected and used for testing.

For the annotations, this study based its approach on the VQA API, introduced on the visualqa website [35]. In this approach, a few sets of requirements need to be met in order to have a model working. First, there is the image step in which images are processed in order to train the model. Next, there is the question step, in which questions need to be input and paired with an image. Lastly, there is the annotation step, in which an answer is introduced and put together with both an image and a question. To facilitate this process, identification numbers (ID) are provided for all questions, answers, and images on the dataset. After all steps, the features, the image, and the question are combined, and the probabilities for each possible answer are assigned.

There are many different ways to utilize the VQA model. The one focused on in this study uses the CNN paired with the BoW. To be able to answer open-ended questions, it is necessary to combine both visual and language understanding. With that in mind, the most common approach to this problem is to use two parts, the first one is to analyze the images (which is the visual aspect), and the second part is to analyze the language (which is the question and answering aspect). The VQA model needs to be able to detect what is being displayed in an image, in order to effectively give an appropriate answer to what is being asked. The CNN part, which will be utilized for analyzing the image, is already discussed in Section 4.

The BoW model utilized in this study is a simple yet widely used approach of representing text data when performing any type of machine-learning application. This model is a representation of the text that describes the occurrence of words within a document. The main part is that it involves a number of known words and a measure of the presence of those words. It has its name because the order of the structure of words inside the document is not important, therefore it can be considered simply to be a “bag” of words [36]. Since this study aims to only use a relatively small dataset (the Sally-UWF dataset) when compared to others, the BoW can be considered to be a great asset. This is because one of the limitations of the BoW is the length of its vocabulary. However, since only dealing with a small, fixed set, where one of the answers will always be the correct one, this model can be very effective. The BoW model turns arbitrary text into fixed-length vectors by counting how many times each word appears. This process is also known as vectorization [37]. For example, looking into our dataset, a few specific sentences were used for vectorization, such as “Is there any damage on this image?” or “How much damage can be seen?”. Looking at these specific questions, a vocabulary can be determined. To create a vocabulary, each word needs to be separated by itself: “Is, there, any, damage, on, this, image, how, much, can, be, seen.” After that, this dataset is vectorized by assigning a number to each word. That means that every time the word appears in a sentence, it is counted. Table 1 depicts an example of this approach.

**Table 1.** An example of the vectorization of a Bag of Words application.

Question	Words											
	is	there	any	damage	in	this	image	how	much	can	be	seen
“Is there any damage in this image?”	1	1	1	1	1	1	1	0	0	0	0	0
“How much damage can be seen?”	0	0	0	1	0	0	0	1	1	1	1	1

In Table 1, vectors of [1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0] and [0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1] are created. After creating a fixed-length vector for each question, the vectors are used as an input for a feedforward neural network. This feedforward neural network takes vectorized inputs, multiplies them by specific weights, and produces an output [37].

For this study, since a relatively simple question dataset is being used, the BoW vectors obtained from our model are used as the input for this feedforward neural network, and passed through two fully connected neural network layers, where every node is connected to every output from the previous layer, to be able to produce an output [37]. Finally, the results of the CNN model and the BoW model are combined and merged together. However, to be able to obtain the conclusive final results for this approach, the “softmax” function was used. This function allows turning the output values into probabilities, in order to quantify the certainty for each answer.

The main focus of the VQA is on the detection accuracy aspect of the proposed pipeline. In total, this model has used 1000 frames and 30 questions, and the questions have 7 possible answers, which are Yes, No, One, Two, Three, Four, More. The further analysis of the VQA and the comparison of the results is presented in Section 6.



## 6. Results and Discussion

In this study, analyses and comparison of all three approaches were conducted [38]. Starting with the Cascade Classifier, the approach successfully detects roof damages at both Martin Hall and Argo Hall, which are the two UWF buildings analyzed in this study. For the analysis, the average detection rate and the average number of false-positive detections are calculated. Figure 5 depicts example results for this approach. Table 2 summarizes the average damage detection rate and average number of false-positive detections using the CC approach. Additionally, the precision and recall values are calculated. The precision is calculated as the number of true positives over the number of true positives and false positives. It is a measure of how valid the predictions made by the model are. False positives reduce the precision of the model. The recall is calculated as the number of true positives over the number of true positives and false negatives. It is a measure of the completeness of the model predictions, so false negatives reduce the recall value. In general, a precision-recall curve is created with different parameters (thresholds) for a single frame, and it allows a common base to compare the different algorithms. With changing threshold, it is expected to have decreased the precision while the recall is increasing. In this study, however, the best parameters are kept fixed, and the precision and recall plot with results from consecutive frames of the video footage is created. The precision-recall variation plots are depicted in Figures 6 and 7, for the Argo Hall video and the Martin Hall video, respectively. The expected results from the precision-recall plots are that the value should be close to 1 in the precision as well as in the recall in most of the frames.



**Figure 5.** Example results obtained using the CC approach.

**Table 2.** Damage detection with the Cascade Classifier.

	Average Detection Rate	Average Number of False-Positive Detections
Argo Hall	41%	0.13
Martin Hall	41%	0.04

According to the results, the proposed CC-based detection algorithm has an average detection rate of 41% for both buildings. However, it has a small false-detection average (0.13), which demonstrates the accuracy of the algorithm. The reason for the average detection rate is around 41% is the motion of the UAV and the gimbal. Each frame is analyzed further individually, and it is realized that if the UAV and/or gimbal has/have a rapid motion, that leads to a motion blur in that frame, which then leads to the performance degradation of the damage detection algorithm. When the UAV hovers over the roof and gimbal has a constant motion, or the UAV has a constant motion and gimbal is fixed, the detection performance increases. To prove that, parts in the video where there is minimum to no motion blur effect were used for further analysis, and the average detection rate results for those parts turned out to be 48% and 55% for Argo Hall and Martin Hall, respectively (Table 3). More importantly, in the results for Martin Hall, there is no false-

positive detection, and there is only one false-positive detection in the entire series of frames in results for Argo Hall, for those selected sections.

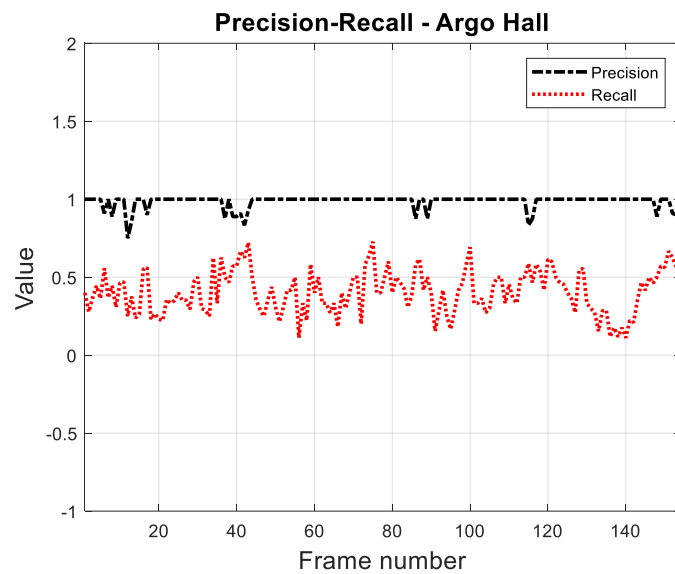


Figure 6. Precision-recall for Argo Hall using the Cascade Classifier.

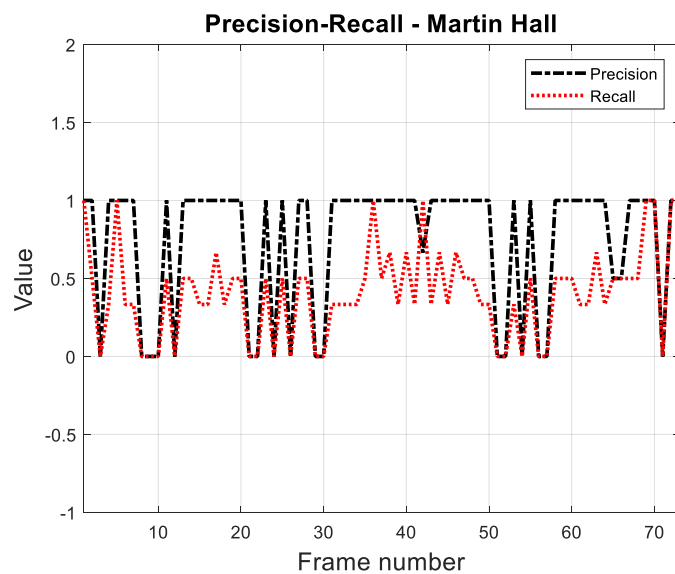


Figure 7. Precision-recall for Martin Hall using the Cascade Classifier.

Table 3. Damage detection with the Cascade Classifier—selected sections.

	Average Detection Rate	Average Number of False-Positive Detections
Argo Hall	48%	0.06
Martin Hall	55%	0.00

The next approach analyzed is CNN. Figure 8 depicts example results for this approach. The average detection rate and the average number of false-positive detection results are given in Table 4, and according to those results, the proposed CNN-based detection algorithm performs better than the CC-based approach. The small false-detection average also shows that the proposed algorithm is accurate. Additionally, the precision and recall variation plots, for both buildings, demonstrate a high level of performance. The same

as it was carried out for the CC, the selected frames with a minimum to no motion blur effect are also analyzed, and the average detection rate results for those parts are 85% and 88% for Argo Hall and Martin Hall, respectively (Table 5). In addition, in results for Argo Hall, there is no false-positive detection; and there is only one false-positive detection in the entire series of frames in results for Martin Hall, for those selected sections. The precision-recall plots are depicted in Figures 9 and 10, for Martin Hall video and Argo Hall video, respectively.



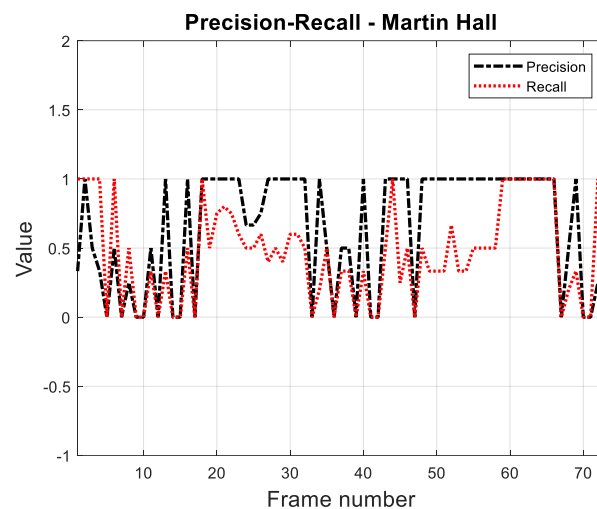
**Figure 8.** Example results using the CNN approach. The images depict frames from aerial footage from Argo Hall and Martin Hall. The CNN detects damages with green squares (light damage) and a yellow square (medium damage), with confidence ratings and the pixel locations.

**Table 4.** Damage detection with CNN.

	Average Detection Rate	Average Number of False-Positive Detections
Argo Hall	55%	0.15
Martin Hall	48%	0.29

**Table 5.** Damage detection with CNN—selected sections.

	Average Detection Rate	Average Number of False-Positive Detections
Argo Hall	85%	0.00
Martin Hall	88%	0.08



**Figure 9.** Precision-recall for Martin Hall using the CNN.

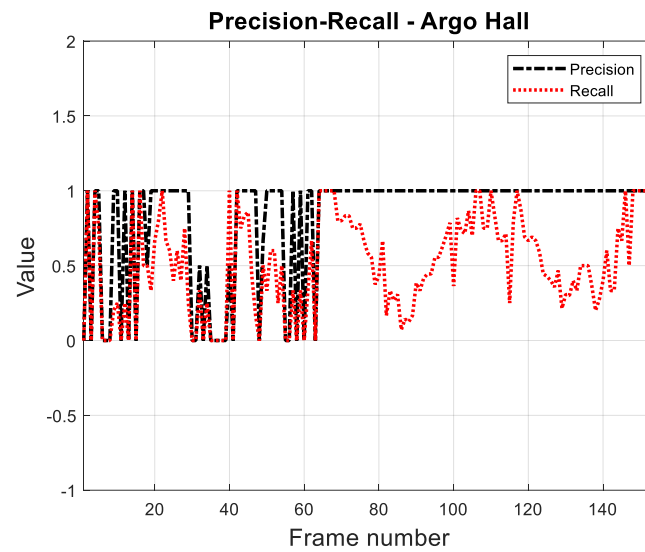


Figure 10. Precision-recall for Argo Hall using the CNN.

The last approach analyzed in this paper is the VQA. Figures 11 and 12 depict a few example results of this approach. As discussed previously, the same corresponding selected frames obtained from two different videos were used. However, differently from previous approaches, in order to identify the damage instances inside each frame, a few different questions were posed to the model. The questions regarding whether there are any damage instances in the image being analyzed, how many buildings were affected, and how many damages can be identified were the main questions used in this study.

Overall, the developed VQA algorithm combination of the CNN and BoW model obtained over 90% accuracy validation (Table 6). Specifically, the overall accuracy of this method considering any type of damage on Argo Hall was 92%, and considering damage occurrences on Martin Hall, it resulted in an overall accuracy of 93%. These results suggest that the developed approach is able to successfully understand and answer questions regarding damage detection. Please note that the high accuracy of this feasibility study is an indication of overfitting, because all images are taken from the UWF campus where the buildings look alike. The lack of diversity in images and the similarity of the training and test dataset inflated the accuracy of the proposed algorithm. However, we expect the accuracy to decrease nominally as we evaluate the method on a more comprehensive dataset.

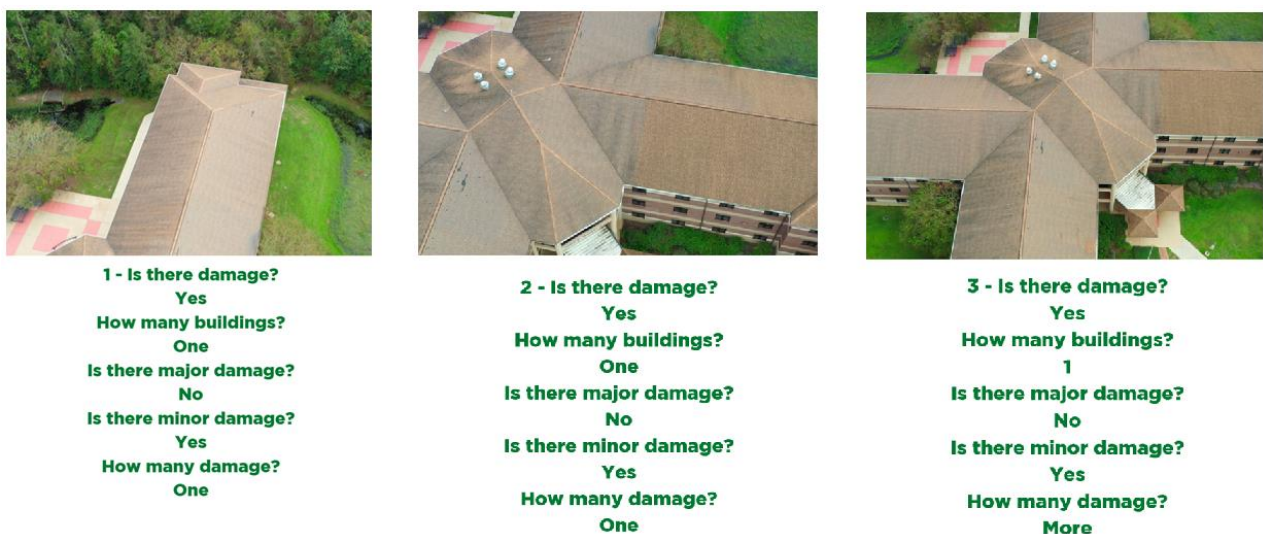


Figure 11. Examples 1–3: VQA Analysis of Post-disaster Footage.



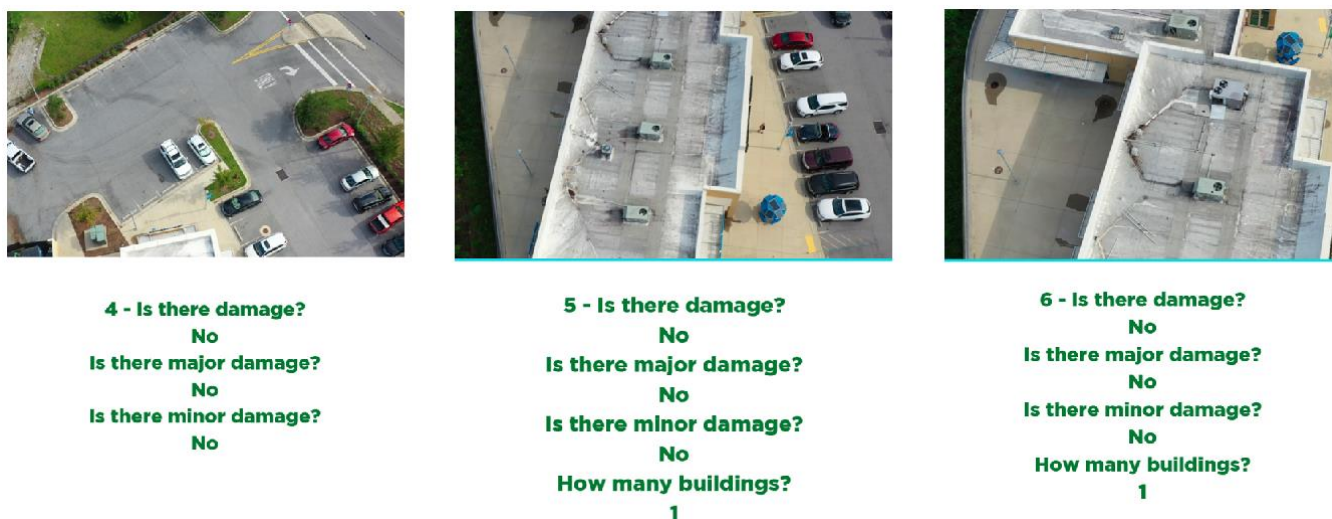


Figure 12. Examples 4–6: VQA analysis of post-disaster footage.

Table 6. Accuracy of VQA Analysis.

	Overall Accuracy	Accuracy for Yes/No	Accuracy for Count
Argo Hall	92%	94%	90%
Martin Hall	93%	96%	89%

According to the results, the approach that had a better performance when trying to detect damages is the VQA; the VQA had an accuracy of 92% and 93%, the CNN had accuracy of 85% and 88%, and the CC had 48% and 55% for the selected region of the test data for Argo Hall and Martin Hall, respectively. The main reason for the CC having a lower accuracy level when compared to other models was about the training process. By acquiring a larger dataset of positive and negative images, as well as increasing the training stages, it is likely that the final accuracy of this CC approach might also significantly increase.

Furthermore, the CNN and VQA approaches had similar results. This is expected because the main reason why the CNN model did not have a relatively higher overall accuracy than VQA was that it failed to detect specific instances of damage due to the motion blur. However, most of this was mitigated in the VQA approach because of the specific questions being asked. Since the VQA model was asked mainly yes/no questions, or whether there were damaged buildings, any specific instance of a single damage being missed is harder, as it would just be counted as part of the whole building.

Additionally, when comparing all approaches, both the CC and CNN provide an easier-to-follow damage detection result, which can prove to be useful in post-disaster scenarios. Both models simply point out all the detected damages. Due to this, when only looking for any instances of damage, these approaches would be better to utilize. However, when looking for more details inside an image, the VQA model can provide more information than the other models, due to the fact that it has a language model paired to the image classification. Analysis shows that the VQA approach has a lot more options when analyzing an image, which proves the feasibility of the use of VQA for post-disaster damage detection.

### 7. Conclusions

In this study, the analysis of three approaches for damage detection, the Cascade Classifier (CC), the Convolutional Neural Network (CNN), and the Visual Question Answering (VQA), were performed. The first model analyzed was the CC, in which an algorithm was developed in order to detect roof damages using aerial footage. Next, the CNN-based detection model was analyzed, and similarly to the CC model, an algorithm to detect

roof damages in aerial footage was developed. Finally, a VQA model was designed by combining an image classification model (CNN) with a language model (Bag of Words), with the goal of not only detecting any damage, but also being able to answer different questions that might arise in post-disaster scenarios. In order to demonstrate their performances, the algorithms of the three approaches were tested on the videos recorded from an Unmanned Aerial Vehicle (UAV) flying over the University of West Florida (UWF) campus after Hurricane Sally in 2020. The results of the overall accuracy performance of each method were presented. The developed VQA algorithm combination of the CNN and BoW model obtained over 90% accuracy validation, specifically, the overall accuracy for Argo Hall was 92%, and the overall accuracy for Martin Hall was 93%, where the CNN had an accuracy of 85% and 88%, and the CC had 48% and 55%, for Argo Hall and Martin Hall, respectively. The VQA analysis results support the feasibility of the use of the developed VQA framework for damage detection in post-disaster aerial footage.

For future work, in order to improve the accuracy, more training data are planned to be used. The custom dataset created for this study, the Sally-UWF dataset, was mainly focused on roof damages. To expand the proposed approach to any type of damage, a larger dataset with different types of damages will be utilized. Furthermore, we plan to expand the experimental validation of data in different regions, as well as compare with other models such as you only look once (YOLO), as another future work.

## 8. Source Code and Data

The source code and data for the VQA approach can be found on our research group's GitHub page at <https://github.com/sevilresearch/VQAdetection>, accessed on 10 April 2023. Additionally, example videos of our post-disaster damage detection research can be found on our research group's YouTube channel at <https://www.youtube.com/watch?v=HW9oE2r65R4> and <https://www.youtube.com/watch?v=0u7murtDRDU>.

**Author Contributions:** Conceptualization, methodology, investigation, writing—original draft preparation, R.D.S.L.; supervision, project administration, writing—review and editing, H.E.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Munich, R.E. Record Hurricane Season and Major Wildfires—The Natural Disaster Figures for 2020. Available online: <https://www.munichre.com/en/company/media-relations/media-information-and-corporate-news/media-information/2021/2020-natural-disasters-balance.html> (accessed on 16 March 2021).
2. Dapena, K. The Rising Costs of Hurricanes. *The Wall Street Journal*, 29 September 2018. Available online: <https://www.wsj.com/articles/the-rising-costs-of-hurricanes-1538222400> (accessed on 16 March 2021).
3. Mohan, A.; Papageorgiou, C.; Poggio, T. Example-based object detection in images by components. *Pattern Anal. Mach. Intell. IEEE Trans.* **2001**, *23*, 349–361. [[CrossRef](#)]
4. Zirakchi, A.; Lundberg, C.; Sevil, H.E. Omni Directional Moving Object Detection and Tracking with Virtual Reality Feedback. In Proceedings of the Dynamic Systems and Control Conference, Tysons Corner, VA, USA, 11–13 October 2017; American Society of Mechanical Engineers: New York, NY, USA, 2017; Volume 58288, p. V002T21A012.
5. Yin, T.; Zhou, X.; Krahenbuhl, P. Center-Based 3D Object Detection and Tracking. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021. [[CrossRef](#)]
6. Viola, P.; Jones, M. Rapid Object Detection Using a Boosted Cascade of Simple Features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001; Volume 1, p. 7176899.
7. Lienhart, R.; Kuranov, A.; Pisarevsky, V. Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection. In *Joint Pattern Recognition Symposium*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 297–304.
8. Wang, X.; Yang, M.; Zhu, S.; Lin, Y. Regionlets for Generic Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 17–24.

9. Zhu, X.; Liang, J.; Hauptmann, A. Msnet: A Multilevel Instance Segmentation Network for Natural Disaster Damage Assessment in Aerial Videos. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 2023–2032.
10. Nex, F.; Duarte, D.; Steenbeek, A.; Kerle, N. Towards real-time building damage mapping with low-cost UAV solutions. *Remote Sens.* **2019**, *11*, 287. [CrossRef]
11. Pi, Y.; Nath, N.; Behzadan, A.H. Disaster impact information retrieval using deep learning object detection in crowdsourced drone footage. *Proc. Int. Workshop Intell. Comput. Eng.* **2020**, 134–143.
12. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
13. Zhou, B.; Tian, Y.; Sukhbaatar, S.; Szlam, A.; Fergus, R. Simple Baseline for Visual Question Answering. *arXiv* **2015**, arXiv:1512.02167.
14. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, L.; Parikh, D. VQA: Visual Question Answering. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
15. Chen, K.; Wang, J.; Chen, L.-C.; Gao, H.; Xu, W.; Nevatia, R. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv* **2015**, arXiv:1511.05960.
16. Shih, K.J.; Singh, S.; Hoiem, D. Where to Look: Focus Regions for Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4613–4621.
17. Yang, Z.; He, X.; Gao, J.; Deng, L.; Smola, A. Stacked Attention Networks for Image Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 21–29.
18. Fukui, A.; Park, D.H.; Yang, D.; Rohrbach, A.; Darrell, T.; Rohrbach, M. Multimodal compact bilinear pooling for visual question answering and visual grounding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016.
19. Kim, J.-H.; On, K.-W.; Lim, W.; Kim, J.; Ha, J.-W.; Zhang, B.-T. Hadamard Product for Low-Rank Bilinear Pooling. *arXiv* **2016**, arXiv:1610.04325.
20. Yu, Z.; Yu, J.; Fan, J.; Tao, D. Multi-Modal Factorized Bilinear Pooling with Coattention Learning for Visual Question Answering. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1821–1830.
21. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical Question-Image Co-Attention for Visual Question Answering. In Proceedings of the NeurIPS, Barcelona, Spain, 5–10 December 2016.
22. Lobry, S.; Marcos, D.; Murray, J.; Tuia, D. RSVQA: Visual question answering for remote sensing data. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8555–8566. [CrossRef]
23. Silva, F.H.; Bezerra, G.M.; Holanda, G.B.; de Souza, J.W.M.; Rego, P.A.; Neto, A.V.L.; de Albuquerque, V.H.C.; Filho, P.P.R. A novel feature extractor for human action recognition in visual question answering. *Pattern Recognit. Lett.* **2021**, *147*, 41–47. [CrossRef]
24. Vo, H.Q.; Phung, T.H.; Ly, N.Q. VQASTO: Visual Question Answering System for Action Surveillance Based on Task Ontology. In Proceedings of the 2020 7th NAFOSTED Conference on Information and Computer Science (NICS), Ho Chi Minh City, Vietnam, 26–27 November 2020; pp. 273–279.
25. Abacha, A.B.; Hasan, S.A.; Datta, V.V.; Liu, J.; Demner-Fushman, D.; Müller, H. VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF. In Proceedings of the CLEF 2019—Conference and Labs of the Evaluation Forum, Lugano, Switzerland, 9–12 September 2019.
26. Hela, V.; Lubna, A.; Kalady, S. CarParkingVQA: Visual Question Answering Application on Car Parking Occupancy Detection. In Proceedings of the 2022 IEEE 19th India Council International Conference (INDICON), Kochi, India, 24–26 November 2022; pp. 1–6.
27. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6999–7019. [CrossRef] [PubMed]
28. Sarkar, A.; Rahnemoonfar, M. VQA-Aid: Visual Question Answering for Post-Disaster Damage Assessment and Analysis. *arXiv* **2021**, arXiv:2106.10548.
29. Sasha, S. A Comprehensive Guide to Convolutional Neural Networks—The eli5 Way. Medium. Towards Data Science. 17 December 2018. Available online: <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53> (accessed on 10 April 2023).
30. Gad, A.F. Faster R-CNN Explained for Object Detection Tasks. Paperspace Blog. 9 April 2021. Available online: <https://blog.paperspace.com/faster-r-cnn-explained-object-detection/> (accessed on 10 April 2023).
31. Chowdhury, T.; Rahnemoonfar, M.; Murphy, R.; Fernandes, O. Comprehensive Semantic Segmentation on High Resolution UAV Imagery for Natural Disaster Damage Assessment. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 3904–3913. [CrossRef]
32. Agrawal, A.; Kembhavi, A.; Batra, D.; Parikh, D. C-vqa: A Compositional Split of the Visual Question Answering (VQA) v1.0 Dataset. *arXiv* **2017**, arXiv:1704.08243.
33. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the v in vqa Matter: Elevating the Role of Image Understanding in Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6904–6913.

34. Kant, Y.; Moudgil, A.; Batra, D.; Parikh, D.; Agrawal, H. Contrast and Classify: Training Robust VQA Models. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 1584–1593. [[CrossRef](#)]
35. Visualqa. VQA: Visual Question Answering. Available online: <https://visualqa.org/download.html> (accessed on 17 January 2022).
36. Jason, B. A Gentle Introduction to the Bag-of-Words Model. Machine Learning Mastery. 7 August 2019. Available online: <https://machinelearningmastery.com/gentle-introduction-bag-words-model/> (accessed on 10 January 2022).
37. Zhou, V. A Simple Explanation of the Bag-of-Words Model. 30 November 2019. Available online: <https://victorzhou.com/blog/bag-of-words/> (accessed on 10 January 2022).
38. Lowande, R. Visual Question Answering (VQA) Analyses for Post-Disaster Damage Detection and Identification Using Aerial Footage. Master’s Thesis, University of West Florida, Pensacola, FL, USA, 2022.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.