Multi-level Attention Network with Weather Suppression for All-weather Action Detection in UAV Rescue Scenarios

Yao Liu[1], Binghao Li[2], Claude Sammut[1], and Lina Yao[3,1]

[1] School of Computer Science and Engineering, University of New South Wales, Sydney, NSW, Australia
[2] School of Minerals and Energy Resources Engineering, University of New South Wales, Sydney, NSW, Australia
[3] Data 61, CSIRO, Sydney, NSW, Australia
{yao.liu3,binghao.li,c.sammut,lina.yao}@unsw.edu.au

Abstract. Unmanned Aerial Vehicles (UAVs) possess significant advantages in terms of mobility and range compared to traditional surveillance cameras. Human action detection from UAV images has the potential to assist in various fields, including search and rescue operations. However, UAV images present challenges such as varying heights, angles, and the presence of small objects. Additionally, they can be affected by adverse illumination and weather conditions. In this paper, we propose a Multi-level Attention network with Weather Suppression for all-weather action detection in UAV rescue scenarios. The Weather Suppression module effectively mitigates the impact of illumination and weather, while the Multi-level Attention module enhances the model's performance in detecting small objects. We conducted detection experiments under both normal and synthetic harsh conditions, and the results demonstrate that our model achieves state-of-the-art performance. Furthermore, a comparison of relevant metrics reveals that our model strikes a balance between size and complexity, making it suitable for deployment on UAV platforms. The conducted ablation experiments also highlight the significant contribution of our proposed modules.

Keywords: Unmanned Aerial Vehicles · Human action detection · Weather suppression · Multi-level attention.

1 INTRODUCTION

With their advantages of high mobility, flexible deployment, and a large surveillance range, Unmanned Aerial Vehicles (UAVs) have progressively demonstrated their utility in surveillance, target tracking, aerial photography, and rescue operations in recent years [40, 8, 52, 3, 2]. Particularly during natural disasters, UAVs can be rapidly deployed to remote areas for swift and extensive scanning, making them highly valuable for search and rescue operations. However, the current application of UAVs in search and rescue often involves capturing aerial

photographs, which are subsequently manually analyzed to identify individuals requiring rescue. This heavy reliance on human intervention makes search and rescue operations involving UAVs labor-intensive. Although object detection has been extensively studied for many years, there are still certain limitations in UAV image detection. UAV images exhibit variations in terms of heights, angles, object scales, and backgrounds, compared to conventional camera images [47]. In the field of search and rescue, UAV images are frequently impacted by inclement weather and challenging illumination conditions. Moreover, UAV rescue operations specifically focus on identifying individuals in need of assistance, requiring the capability to differentiate between various human actions to determine the urgency of rescue. The differentiation of human actions helps exclude unrelated individuals and expedite the search and rescue process. However, the subtle variations between human actions pose a greater challenge for distinction compared to the distinctions between categories in general object detection.

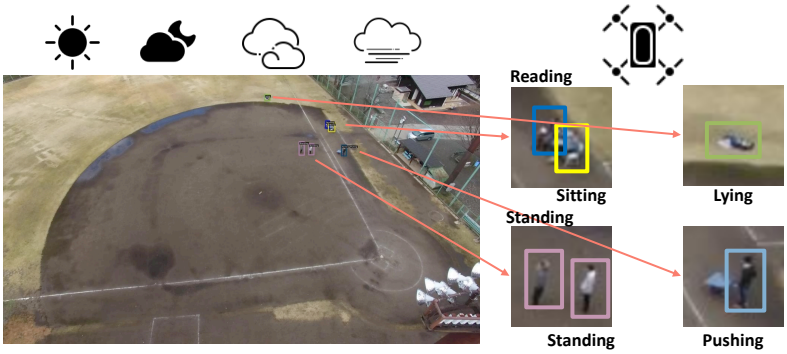

Fig. 1. Human action detection on UAV images. In different illumination and weather conditions, we perform action recognition and localization of humans in UAV-captured images. Strong lighting and nighttime conditions can cause objects to lose textural detail and blend in with the background. Adverse weather conditions can obscure parts of the object and reduce overall visibility.

Apart from manual recognition of UAV images, traditional machine learning approaches often employ a sliding window paradigm and rely on hand-crafted features [28, 46]. However, these methods are time-consuming, and their feature robustness is insufficient. In recent years, the advancements in deep learning have led to the emergence of Convolutional Neural Networks (CNNs) [34] and Generative Adversarial Networks (GANs) [14]. These techniques have significantly impacted the field of object detection and have naturally extended their influence to UAV image detection as well. UAV image detection encompasses various tasks, such as object detection [53], dense detection [30], and object counting [5]. However, it currently encounters challenges related to small object detection [20] and handling long-tailed distribution of objects [51]. Hence, UAV image detec-

tion cannot be directly adopted from conventional object detection methods, and further improvements are necessary to address its unique challenges and requirements. Indeed, UAV images present unique tasks, especially in human detection, which encompass gesture detection [23] and action detection [1]. In the context of search and rescue, UAVs face the challenge of operating under complex conditions and conducting extensive scanning operations. Due to their inherent limitations, UAVs often cannot maintain steady recording of small areas, resulting in a shorter duration for capturing human activity compared to regular circumstances. Detection methods that rely on video analysis typically require multiple consecutive frames, preferably with a stable background [26, 31]. Consequently, detecting humans and recognizing their actions in search and rescue scenarios using a single image poses a significant and practical problem [27].

Our research focuses on human action detection in UAV images under various illumination and weather conditions, as illustrated in Figure 1. In our study, we utilize UAV data that encompasses diverse angles, altitudes, and backgrounds. Additionally, we incorporate a wide range of challenging conditions, including strong lighting, nighttime scenarios, cloudy weather, and foggy environments. Our objective is to create a human action detection model that can effectively operate in various scenarios using just a single image. This capability is particularly valuable in demanding environments like areas affected by natural disasters or underground mines, where video-based methods may not be feasible or practical. Our model consists of two primary modules: the Weather Suppression module and the Multi-level Attention module. The Weather Suppression module is designed to extract the noise map and illumination map from the backbone network, enabling the acquisition of True Color map features. This separation process helps suppress weather-related noise and improve the quality of the image features. Furthermore, the Weather Suppression module can be seamlessly integrated into the model's neck, forming an end-to-end network. This integration is distinct from the approach of training separate image enhancement networks independently. To achieve more effective cross-level feature aggregation, we employ the Multi-level Attention module. This module intelligently aggregates multi-level features by assigning weights based on the attention mechanism. This enables a rational and adaptive fusion of features from different levels, resulting in improved performance and better representation of the target objects. Through our experiments, we have demonstrated that our model achieves state-of-the-art performance in this task.

Our main contributions are as follows:

- We introduce a novel Multi-level Attention network with Weather Suppression specifically designed for all-weather action detection in UAV rescue scenarios. Additionally, we synthesize a multi-weather human action detection dataset by augmenting an existing dataset, facilitating effective training and accurate detection under various weather conditions encountered during UAV rescue missions.
- The Weather Suppression module effectively mitigates the adverse effects caused by weather and illumination, resulting in detection results that closely

resemble those achieved under normal weather conditions. By separating the noise map and illumination map from the backbone network, we enhance the quality of image features, thus improving the robustness of our model.
- The Multi-level Attention module dynamically adjusts the weights using the attention mechanism after the Feature Pyramid Network (FPN), facilitating effective cross-level feature aggregation. This module improves the model's ability to focus on relevant features and adaptively fuse them, enhancing the overall detection capability.
- Our model excels in human action detection in UAV images. Through comprehensive comparison experiments and ablation studies, we demonstrate the superior performance of our model, highlighting the significant contributions made by the Weather Suppression and Multi-level Attention modules.

2 RELATED WORK

2.1 Object detection

In the past decade, object detection has made significant progress, with the mainstream detectors being divided into two-stage and single-stage approaches. Single-stage detectors are generally faster but slightly less accurate compared to two-stage detectors. R-CNN [13] is a milestone in introducing deep learning methods to the field of object detection, and it belongs to the two-stage method. The subsequent advancements include Fast R-CNN [12], which integrates the classification head and regression head into the network, and Faster R-CNN [38], which proposes the Region Proposal Network (RPN) to generate region proposals, forming an end-to-end object detection framework. Many of the later two-stage methods are built upon the improvements made by Faster R-CNN. For example, Mask R-CNN [15] reduces the accuracy loss of ROI pooling by introducing ROI-align. The pioneering work on single-stage detectors is YOLO [36], which has led to the development of numerous detectors in the YOLO series, including the widely used YOLOv3 [37] and the latest YOLOx [10]. Additionally, there are other mainstream basic methods, including SSD [25] and RetinaNet [21]. However, after YOLOv3, the integration of Anchor Box, Focal Loss, and Feature Pyramid Network (FPN) gradually became more prevalent. Due to the speed and increased accuracy offered by single-stage detectors, UAV image detection has mostly focused on improving single-stage detectors [19].

In the context of the detection task, the model neck plays a critical role as it serves as the connection between the backbone and the detection head. The backbone is responsible for extracting feature maps, often using the classification field's outputs directly. On the other hand, the detection head typically consists of both classification and regression modules. Therefore, the model neck, which aggregates feature map information, is crucial for effective object detection. Initially, methods like SSD [25] utilized the multi-level output of the backbone for detection but lacked explicit feature aggregation, leading to the evolution of the model neck. Subsequently, the Feature Pyramid Network (FPN) emerged as the mainstream approach for model necks, and it has been adopted by many models

such as YOLOv3 [37], RetinaNet [21], and Faster R-CNN [38]. FPN utilizes a top-down aggregation process to enhance feature maps. Later, PANet [6] introduced bottom-up aggregation to further improve performance. Building upon these advancements, subsequent methods introduced additional techniques to refine feature fusion in the model neck. For example, ASFF [24] employed an attention mechanism to determine fusion weights, while NAS-FPN [11] and BiFPN [44] focused on identifying important blocks for repeated fusion. In a similar vein, the study mentioned in [20] leveraged average pooling and deconvolution to enhance FPN performance, particularly for detecting small objects in UAV images. These developments highlight the continuous pursuit of refining the model neck to improve object detection performance.

2.2 UAV image detection

The flight height of UAVs can vary from a dozen to hundreds of meters, and the cameras mounted on gimbals can capture images at different angles. As a result, UAV images contain objects with large scale variations and often include side and top views, making them challenging for conventional object detection methods. Fortunately, UAV images typically have fewer overlapping objects due to the shooting angle, reducing the need to focus extensively on handling overlapping objects [19]. Earlier research in UAV image detection often involved adapting general object detection methods to UAV settings [41, 42]. Popular UAV datasets, such as VisDrone [54], UAVDT [7], and UAV123 [29], provide labeled data for various object classes, enabling common tracking and detection tasks. To address the specific challenges of UAV image detection, researchers have made improvements to single-stage detectors based on methods like SSD and YOLO. For example, [39, 4] enhance the SSD method, while [35, 45] modify the YOLO method to make single-stage detectors more suitable for UAV image detection. The UAV-Gesture dataset [32] contains a variety of drone command signals, but it focuses on close-range interactions between drones and people, with low-altitude flights and minimal background changes, resembling typical surveillance videos rather than distinct UAV image scenarios. On the other hand, the Okutama-Action dataset [1] consists of drone videos that specifically capture human actions, encompassing various heights, angles, and backgrounds. [27] utilizes the Okutama-Action dataset and their own dataset to apply UAV image human action detection to real-world rescue operations, demonstrating the practical implications of their research.

2.3 Weather impact

When operating outdoors, especially during search and rescue missions, UAVs often encounter challenging environmental conditions such as unfavorable illumination (e.g., strong light or nighttime) and adverse weather (e.g., cloudy or foggy conditions). These factors can have a significant impact on the visibility and appearance of objects captured in UAV images, leading to less accurate detection results [47]. Strong light and nighttime conditions can cause objects to

lose texture details, undergo color distortions, and blend into the background. Unfavorable weather conditions such as clouds and fog can further obscure parts of objects, making them appear incomplete and reducing overall visibility. These factors pose challenges to accurate detection in UAV images. While illumination variations and weather conditions also affect image quality in general object detection, they become critical factors in UAV image detection, particularly in response to special events where UAVs play a crucial role in search and rescue operations [18]. Consequently, suppressing the adverse effects of unfavorable illumination and weather can improve the overall image quality and, in turn, enhance the performance of object detection algorithms in UAV images [17, 50].

3 METHOD

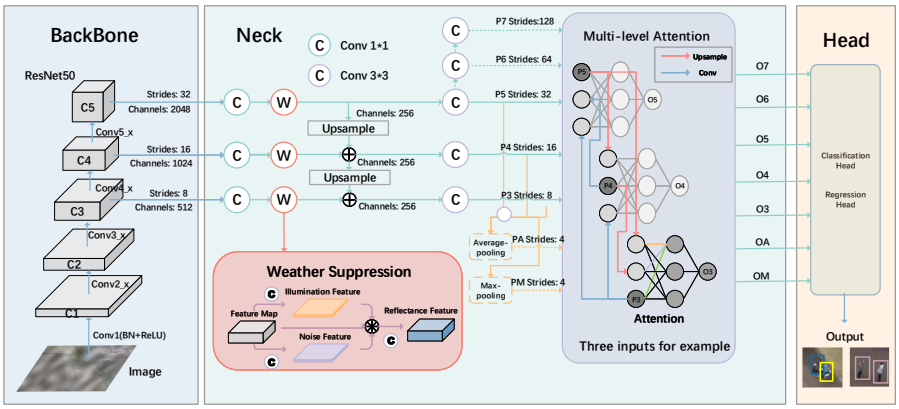

Fig. 2. Overview of our model. The model backbone is ResNet50 with multi-level output, the neck contains a Weather Suppression module and a Multi-level Attention module, and the head includes the Focal Loss.

3.1 Overview

Our model is designed to perform human action detection on UAV images. It takes a single image as input and provides the action category and bounding box for each person as output. The overview of our model is illustrated in Figure 2. To strike a balance between performance and speed, we employ a medium-sized ResNet50 [16] as the backbone network. Following the approach of YOLO [36], SSD [25], and RetinaNet series, we utilize multi-level output. To optimize computational efficiency, we output C3-C5 features into the model neck. Within the model neck, we incorporate two key modules: the Weather Suppression module

and the Multi-level Attention module. The Weather Suppression module aims to minimize the influence of illumination and weather conditions on the image. On the other hand, the Multi-level Attention module intelligently assigns weights for multi-level feature synthesis, promoting effective feature aggregation. For the model head, we adopt a RetinaNet-like classification head and regression head. To optimize the learning process, we utilize the Focal Loss [21], which reinforces the learning of challenging samples. This is particularly important for human action detection, as the differences between various actions can be subtle.

3.2 Weather Suppression

According to the Retinex theory [17, 50], an image can be decomposed into a reflectance map and an illumination map. The reflectance map represents the inherent properties of the objects in the image and remains relatively constant, while the illumination map captures the variations in light intensity. In our model, we consider the original image as $I_o \in \mathbb{R}^{w \times h \times 3}$, where w and h represent the width and height of the image, respectively. This image can be decomposed into the reflectance map $I_r \in \mathbb{R}^{w \times h \times 3}$ and the illumination map $I_i \in \mathbb{R}^{w \times h \times 1}$. This separation is performed by element-wise pixel multiplication, denoted as $\odot$. By utilizing the reflectance map, we can capture the true colors of the objects in the image and enhance the robustness of computer vision tasks.

$$I_o = I_r \odot I_i \tag{1}$$

In addition to the impact of illumination, UAV images are also influenced by various weather effects. Cloudy and foggy conditions, in particular, can introduce noise that affects the details of the image and hampers the detection of small objects. To address this issue, we incorporate noise factors $N \in \mathbb{R}^{w \times h \times 1}$ into the composition of the original image, resulting in a final equation. The addition of noise factors to the original image formulation is a key step in our model, as it helps to enhance the robustness and accuracy of object detection in challenging weather conditions.

$$I_o = I_r \odot I_i + N \tag{2}$$

Thus, we can obtain the reflection map from the original image using a simple operation. However, it is important to note that to avoid encountering division by zero in the illumination map, we define $\tilde{I}_i = \frac{1}{I_i}$.

$$I_r = (I_o - N) \odot \tilde{I}_i \tag{3}$$

Unlike the approach taken in Darklighter [50], where the original images are processed separately, we adopt a integrated strategy to handle the adverse effects of illumination and weather. Separately processing the original images would require a extra network for iterative optimization, which cannot be easily integrated into an end-to-end framework with subsequent tasks. In our model, we aim to address these challenges of illumination and weather simultaneously

within the detection task. To achieve this, we feed the images directly into the backbone network for feature extraction, similar to regular detection tasks. In this way, the feature maps obtained after the backbone contain not only the features of the reflection map but also the features of the illumination map and noise. Therefore, we perform the weather suppression process after the backbone. As depicted in Figure 2, for each level of feature maps that require further processing, we employ the Weather Suppression module. This module is responsible for separating the features of the illumination map and noise, and then calculating the features of the reflection map for subsequent processing. It is important to note that the Weather Suppression module of each level does not share parameters.

3.3 Multi-level Attention

In our own efforts, we have developed a multi-level attention mechanism. Our model neck incorporates the attention module after the Weather Suppression and FPN modules. This attention module learns the importance of each level and assigns fusion weights more rationally. Unlike top-down, bottom-up, or search-based approaches, the weights determined by the attention module are based on network learning, making them more reasonable and effective. By combining top-down and bottom-up feature flows, our model gains a better understanding of the contextual relationships within an image. The top-down pathway allows the model to incorporate high-level semantic information, while the bottom-up pathway captures fine-grained details. This contextual awareness enables the model to make more informed decisions about object detection. The attention mechanism we've introduced facilitates the adaptive fusion of features from different levels. This is a crucial advantage, as it allows the model to dynamically adjust the contribution of each feature level based on the specific requirements of the detection task. This adaptability ensures that the model allocates more resources to relevant levels, leading to improved performance.

Given our specific task of human action detection in UAV images, which involves relatively small objects and considerations for network size, we choose to perform fusion at the third to fifth levels. Human actions in UAV images can vary significantly in scale due to differences in altitude, angles, and distances. The combination of top-down and bottom-up feature propagation helps our model handle these scale variations effectively. The multi-level attention mechanism contributes to the model's ability to generalize well across different scenarios. By adaptively selecting and fusing features, the model can learn to extract relevant information regardless of changes in lighting, weather conditions, or object sizes. This enhanced generalization is particularly valuable for real-world applications where UAVs may encounter diverse environments.

$$O3 = Attn(P3, Upsample(P4), Upsample(P5)) \tag{4}$$

$$O4 = Attn(Conv(P3), P4, Upsample(P5)) \tag{5}$$

$$O5 = Attn(Conv(P3), Conv(P4), P5) \tag{6}$$

Certainly, our model includes deeper networks for feature fusion from level 3 to level 7, as well as Max-pooling and Average-pooling operations. These components will be discussed and evaluated in our ablation experiments to assess their impact on the model's performance.

4 EXPERIMENT

4.1 Dataset

We conducted our experiments using the Okutama-action dataset [1]. The dataset was captured by UAVs at a resolution of 4k and a frame rate of 30 FPS. The data was collected at various flight heights ranging from 10 to 45 meters, with the camera angle set at either 45 degrees or 90 degrees. The dataset consists of 22 scenes, each involving up to 9 actors.

For our baseline human action detection, we used a subset of the dataset that contains 77,365 frames of images with a resolution of 1280x720. Among these frames, 60,039 were used for training and 17,326 for testing. The dataset includes 12 common actions, namely Handshaking, Hugging, Reading, Drinking, Pushing/Pulling, Carrying, Calling, Running, Walking, Lying, Sitting, and Standing. Additionally, the dataset also includes instances of humans with undefined actions, which we refer to as "Others", following the approach in [27]. Hence, in total, we have 13 action categories. However, the presence of multiple simultaneous actions in the dataset complicates the classification task. To address this, we reclassify the action labels into atomic categories: Running, Walking, Lying, Sitting, and Standing. This categorization is similar to the six categories proposed in [27], with the addition of Waving in their case. We exclude Waving as a separate category since it can still be performed alongside other actions. As a result, we eliminate all action labels except for a single label, "Person", which represents human detection. Finally, we conducted experiments on the Okutama-action dataset using three different label divisions: the original 13 categories, the reclassified 5 categories, and the single category (Person) only. These divisions are depicted in Figure 3.

4.2 Pre-processing

The images in the Okutama-action dataset are frames extracted from videos. To enhance efficiency and reduce redundancy, we adopt the common practice of selecting one frame every ten frames for training purposes. Most networks used in our experiments resize the input images to a range of 300 to 600 pixels. Compared to other datasets like MS COCO [22] and Pascal VOC [9], the Okutama-action dataset has higher resolution images with smaller objects. Resizing the images using a standard method would result in severe compression of object pixels, leading to incorrect detection. To overcome this issue, we employ a segmentation approach as a pre-processing step. We divide a 1280x720 pixel image into six segments of size 640x480, with a 50% overlap in both horizontal and vertical

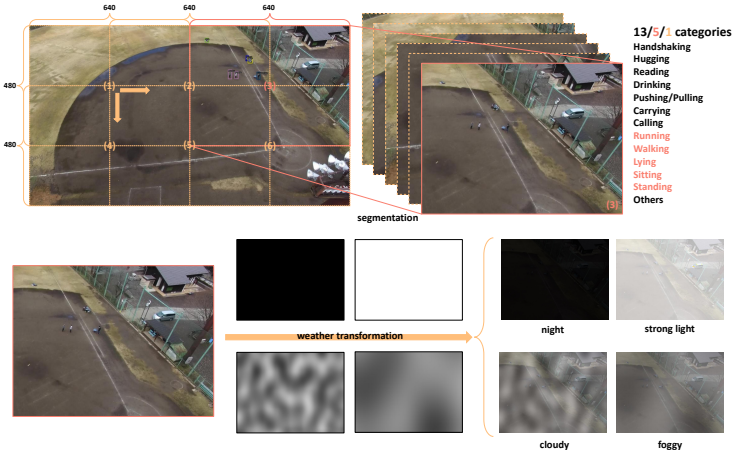

Fig. 3. Okutama-action dataset and data pre-processing.

directions. This results in a 3x2 grid segmentation, as illustrated in Figure 3. This segmentation strategy offers several advantages. First, it serves as a data augmentation technique by providing a larger training dataset, which helps the deep network in learning discriminative features. Second, segmentation preserves the details of the objects without compression, taking advantage of the high-resolution nature of the images and facilitating the detection of small objects. Lastly, the 50% overlap ensures that objects are not split across different sub-images, reducing the risk of missing detections due to cutoff boundaries.

The Okutama-action dataset covers scenes with varying weather conditions (sunny and cloudy) and illumination conditions (morning and afternoon). However, these conditions may not include extreme disturbances encountered in challenging environments like search and rescue scenarios. To address this limitation and simulate harsh conditions, we employ a method inspired by [5] for image transformation. Specifically, we perform pixel-level transformations to synthesize four different severe conditions: extreme illumination (strong light and night) and unfavorable weather (cloudy and foggy). The image transformation follows an image blending algorithm [43], and the process is described in Equation 7. This transformation allows us to augment the dataset with images that capture the challenges posed by extreme lighting and adverse weather conditions, enabling the model to learn robustness and improve its performance in real-world scenarios.

$$\Phi = \alpha I + (1 - \alpha)N + P \tag{7}$$

The image transformation process involves several components: the original image I, the noise map N for simulating different illumination and weather conditions, the perturbation factor P for brightness correction, the weight ratio α, and the resulting transformed image Φ. The noise maps N are categorized

into four types. For illumination transformations, we use all-0 or all-255 matrices to represent strong light and night conditions. For weather transformations, we generate simulated clouds or fog using Perlin noise [33], as shown in Figure 3. The perturbation factor P is set as an integer multiple of the all-1 matrix, which is used to correct the overall brightness of the image. The weight ratio α is set to 0.3, highlighting the impact of harsh weather during the transformation process. During training and testing, the experiments are divided into two groups: the first group does not undergo any weather transformation, while the second group undergoes weather transformation. In the second group, each image is randomly transformed using one type of noise, providing variability and diversity in the dataset.

4.3 Experiments

The evaluation metrics used for assessing the performance of the human action detector in our experiments are based on the mean Average Precision (mAP), which is also commonly used in the MS COCO dataset evaluation [22]. To calculate the mAP, we use the Intersection over Union (IoU) metric, which measures the overlap between the predicted bounding box and the ground truth box. Specifically, we choose an IoU threshold of 0.5, which is a common practice in evaluating UAV image detectors [1, 27, 48, 49].

Our standard model employs features from P3 to P5, with a weather suppression module for adverse weather conditions. Models for our experiment comparison are as follows. The data pre-processing part is consistent for all reproduction models.

- Okutama-action [1]: These results are obtained from the original Okutama-action paper. The method used is similar to SSD, and the experiments are conducted using both 12 categories and 1 category. The results for the 12 categories can be roughly compared with our 13 categories.
- Drone-surveillance [27]: This paper focuses on using UAV images in the rescue field and utilizes both the Okutama-action dataset and their own dataset. For the Okutama-action dataset, they conduct experiments using both 6 categories and 1 category. The results for the 6 categories can be roughly compared with our 5 categories.
- DIF R-CNN [48]: This method is a pedestrian detection approach based on Faster R-CNN with context information enhancement. The authors initially conducted experiments using 1 category of the Okutama-action dataset, but the experiment setup was not consistent with the standard split of the training and testing sets. In a subsequent paper [49], they updated the experiment and provided revised results using the same method, which we cite for comparison.
- TDFA [49]: This method is a two-stream video-based detection approach. The best results on the Okutama-action dataset require a total of 9 frames of input images. Since the video-based method cannot be directly compared with other methods, we consider its experiments with single-image input separately.

- SSD [25]: We reproduce the standard SSD method on the Okutama-action dataset, using 512-pixel input resolution, which has shown improved results.
- RetinaNet [21]: We reproduce the RetinaNet model, using ResNet as the backbone.
- YOLOv3 [37]: We reproduce the popular YOLOv3 method, utilizing DarkNet53 as the backbone and 608-pixel input resolution.
- YOLOx [10]: We reproduce the state-of-the-art YOLOx model, focusing on the S-mode variant with memory consumption comparable to the other models.

Table 1. Comparison results of human action detection experiments on Okutama-action dataset. * indicates a rough comparison.

mAP@0.5	1 category	5 categories	13 categories
Okutama-action [1]	72.3%	–	18.8%*
Drone-surveillance [27]	–	35.0%*	–
DIF R-CNN [48]	84.1%	–	–
TDFA [49]	78.7%	–	–
SSD [25]	80.3%	39.5%	20.5%
RetinaNet [21]	85.5%	40.6%	20.7%
YOLOv3 [37]	78.3%	35.8%	17.8%
YOLOx [10]	84.3%	36.4%	19.4%
Ours w/o Weather Suppression	**85.7%**	**43.8%**	**23.9%**

Table 2. Comparison results of human action detection experiments on Okutama-action dataset with weather transformation.

mAP@0.5	1 category	5 categories	13 categories
SSD [25]	80.0%	39.4%	20.3%
RetinaNet [21]	85.4%	41.4%	22.8%
YOLOv3 [37]	77.1%	33.6%	19.7%
YOLOx [10]	82.3%	35.4%	22.0%
Ours w/o Weather Suppression	85.4%	41.0%	23.2%
Ours w/ Weather Suppression	**85.6%**	**45.2%**	**24.7%**

In Table 1, we can see that our model, which includes the Multi-level Attention module, achieves state-of-the-art results in all three divisions (1 category, 5 categories, and 13 categories). The second-best results are obtained by RetinaNet. Our model outperforms the second-place model by 0.2% in the 1 category division, and by 3.2% in both the 5 and 13 categories divisions.

In Table 2, we present the results of the models under harsh illumination and weather conditions, including both with and without the Weather Suppression module. In the 1 category division, the detection results of each model are slightly reduced, indicating that harsh conditions have a small impact on the detection results. Our Weather Suppression module only reduces the detection results by 0.1% compared to normal conditions, and the model without the Weather Suppression module still outperforms the other models.

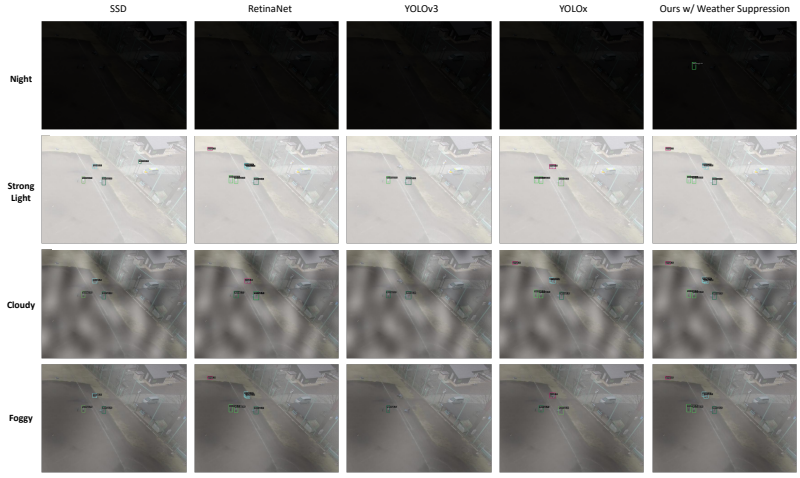

Fig. 4. Comparison of model detection results under harsh conditions.[1]

In the 5 categories division, the detection results of each model decrease significantly, and our model without the Weather Suppression module experiences a decrease of 2.8%. However, with the addition of the Weather Suppression module, our model achieves 45.2% results, surpassing all other models and even exceeding the results under normal conditions.

It is worth noting that in the 13 categories division, some models perform better under unfavorable conditions. This is due to the label definition of the Okutama-action dataset, where the 13 categories include the simultaneous execution of multiple actions. The labels are chosen for the least obvious action categories in order to maintain the highest classification diversity. For example, if a person is sitting and reading, the label assigned is "reading", resulting in a wrong detection result for "sitting". This label definition leads to some misunderstandings in the detection results because the network detects the most obvious categories.

Overall, the atomic classification of 5 categories is more persuasive. Our model performs competitively in all divisions and shows promising results, especially with the inclusion of the Weather Suppression module, which helps mitigate the impact of harsh illumination and weather conditions on the detection performance.

The 5 categories atomic division is considered to have practical significance, and it is visually demonstrated in Figure 4. Notably, only our model achieves correct detections in the night condition, indicating the contribution of the Weather Suppression module. Our model demonstrates advantages under various other harsh conditions as well. The confusion matrix in Figure 5 reveals that Walking,

[1] More comparison results in https://youtu.be/Os9RCcDWgz4.

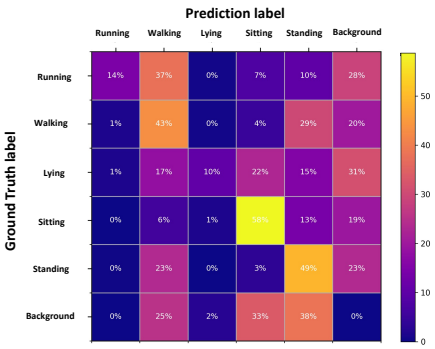

Fig. 5. Confusion matrix for 5 categories.

Table 3. Model Comparison.

	FPS	GFLOPs	Params(M)	Memory(G)
DIF R-CNN [48]	4.55	–	–	–
TDFA [49]	12.50	–	–	–
SSD [25]	17.24	102.80	24.39	11.99
RetinaNet [21]	12.50	61.32	36.10	1.40
YOLOv3 [37]	18.52	58.15	61.52	6.96
YOLOx [10]	16.13	9.99	8.94	6.24
Ours w/o Weather Suppression	11.77	61.59	31.40	2.56
Ours w/ Weather Suppression	11.76	72.74	36.71	2.75

Running, and Standing remain the most challenging categories to classify due to their similar features. This aligns with our perception that these actions are harder to distinguish based on a single image. Lying has distinct features but is more difficult to detect, while Sitting exhibits the best distinguishing features.

Table 3 presents a comparison of several model values. Our model achieves a more balanced situation across various measures while still achieving state-of-the-art results. In terms of complexity, YOLOx has the best performance but requires higher memory usage. YOLOv3 and SSD also demand high memory usage. RetinaNet is the closest in complexity to our model, while DIF R-CNN is slower due to its two-stage detection. The balanced nature of our model makes it well-suited for deployment on UAV platforms with limited hardware resources.

Table 4. Ablation experiments.

mAP@0.5	1 category	5 categories	13 categories	settings
1	83.6%	41.6%	18.4%	P3-P7
2	83.9%	39.5%	19.3%	P3-P5&PA
3	82.6%	38.2%	20.2%	P3-P5&PM
4	85.7%	43.8%	23.9%	P3-P5

In the ablation experiments presented in Table 4, we explore different feature fusion strategies in the attention module. We first investigate the inclusion of higher-level features such as P6 and P7 but find that they do not yield better re-

sults. This can be attributed to the fact that objects in UAV images are smaller in size, and the higher-level features may overlook these small objects. Next, we incorporate Average-pooling and Max-pooling in the attention module, but again, the results do not improve significantly. While pooling features have been shown to enhance model performance in some methods, in our model, attention plays a crucial role in rationalizing feature fusion weights. The inclusion of pooling features can disrupt this fusion process. Based on these findings, our final adopted model utilizes P3-P5 as feature inputs in the attention module, which yields the best results.

5 CONCLUSIONS

In summary, our proposed Multi-level Attention network with Weather Suppression addresses the challenges of all-weather action detection in UAV rescue scenarios. Through rigorous experiments, we demonstrate the effectiveness of our approach in mitigating the effects of illumination and weather conditions, improving model performance, and achieving state-of-the-art results. The balanced performance of our model across various metrics makes it a promising solution for deployment on UAV platforms in search and rescue applications.

References

1. Barekatain, M., Martí, M., et al: Okutama-action: An aerial view video dataset for concurrent human action detection. In: CVPR Workshops. pp. 2153–2160 (2017)
2. Bozcan, I., Kayacan, E.: AU-AIR: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In: ICRA. pp. 8504–8510 (2020)
3. Bozcan, I., Kayacan, E.: Context-dependent anomaly detection for low altitude traffic surveillance. CoRR p. abs/2104.06781 (2021)
4. Budiharto, W., Gunawan, A.A.S., et al: Fast object detection for quadcopter drone using deep learning. In: ICCCS. pp. 192–195 (2018)
5. Cai, Y., Du, D., et al: Guided attention network for object detection and counting on drones. In: MM. pp. 709–717 (2020)
6. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: CVPR. pp. 6154–6162 (2018)
7. Du, D., Qi, Y., et al: The unmanned aerial vehicle benchmark: Object detection and tracking. In: ECCV. pp. 375–391 (2018)
8. Erdelj, M., Natalizio, E.: Uav-assisted disaster management: Applications and open issues. In: ICNC. pp. 1–5 (2016)
9. Everingham, M., Gool, L.V., et al: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. pp. 303–338 (2010)
10. Ge, Z., Liu, S., et al: YOLOX: exceeding YOLO series in 2021. CoRR p. abs/2107.08430 (2021)
11. Ghiasi, G., Lin, T., Le, Q.V.: NAS-FPN: learning scalable feature pyramid architecture for object detection. In: CVPR. pp. 7036–7045 (2019)
12. Girshick, R.B.: Fast R-CNN. In: ICCV. pp. 1440–1448 (2015)
13. Girshick, R.B., Donahue, J., et al: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. pp. 580–587 (2014)

14. Goodfellow, I.J., Pouget-Abadie, J., et al: Generative adversarial networks. Commun. ACM pp. 139–144 (2020)
15. He, K., Gkioxari, G., et al: Mask R-CNN. In: ICCV. pp. 2980–2988 (2017)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
17. Land, E.H.: The retinex theory of color vision. Scientific american pp. 108–129 (1977)
18. Li, T., Liu, J., et al: Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In: CVPR. pp. 16266–16275 (2021)
19. Li, Z., Liu, X., et al: A lightweight multi-scale aggregated model for detecting aerial images captured by uavs. J. Vis. Commun. Image Represent. p. 103058 (2021)
20. Liang, X., Zhang, J., et al: Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis. IEEE Trans. Circuits Syst. Video Technol. pp. 1758–1770 (2020)
21. Lin, T., Goyal, P., et al: Focal loss for dense object detection. In: ICCV. pp. 2999–3007 (2017)
22. Lin, T., Maire, M., et al: Microsoft COCO: common objects in context. In: ECCV. pp. 740–755 (2014)
23. Liu, C., Szirányi, T.: Real-time human detection and gesture recognition for on-board UAV rescue. Sensors p. 2180 (2021)
24. Liu, S., Huang, D., Wang, Y.: Learning spatial fusion for single-shot object detection. CoRR p. abs/1911.09516 (2019)
25. Liu, W., Anguelov, D., et al: SSD: single shot multibox detector. In: ECCV. pp. 21–37 (2016)
26. Mabrouk, A.B., Zagrouba, E.: Abnormal behavior recognition for intelligent video surveillance systems: A review. Expert Syst. Appl. pp. 480–491 (2018)
27. Mishra, B., Garg, D., et al: Drone-surveillance for search and rescue in natural disaster. Comput. Commun. pp. 1–10 (2020)
28. Moranduzzo, T., Melgani, F.: Detecting cars in UAV images with a catalog-based approach. IEEE Trans. Geosci. Remote. Sens. pp. 6356–6367 (2014)
29. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for UAV tracking. In: ECCV. pp. 445–461 (2016)
30. Papaioannidis, C., Mademlis, I., Pitas, I.: Autonomous UAV safety by visual human crowd detection using multi-task deep neural networks. In: ICRA. pp. 11074–11080 (2021)
31. Perera, A.G., Law, Y.W., Chahl, J.: Drone-action: An outdoor recorded drone video dataset for action recognition. Drones (2019)
32. Perera, A.G., Law, Y.W., Chahl, J.S.: UAV-GESTURE: A dataset for UAV control and gesture recognition. In: ECCV. pp. 117–128 (2018)
33. Perlin, K.: Improving noise. ACM Trans. Graph. p. 681–682 (jul 2002)
34. Radovic, M., Adarkwa, O., Wang, Q.: Object recognition in aerial images using convolutional neural networks. J. Imaging p. 21 (2017)
35. Radovic, M., Adarkwa, O., Wang, Q.: Object recognition in aerial images using convolutional neural networks. J. Imaging p. 21 (2017)
36. Redmon, J., Divvala, S.K., et al: You only look once: Unified, real-time object detection. In: CVPR. pp. 779–788 (2016)
37. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. CoRR p. abs/1804.02767 (2018)
38. Ren, S., He, K., et al: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015)

39. Rohan, A., Rabah, M., Kim, S.H.: Convolutional neural network-based real-time object detection and tracking for parrot ar drone 2. IEEE Access pp. 69575–69584 (2019)
40. Semsch, E., Jakob, M., et al: Autonomous UAV surveillance in complex urban environments. In: IAT. pp. 82–85 (2009)
41. Sevo, I., Avramovic, A.: Convolutional neural network based automatic object detection on aerial images. IEEE Geosci. Remote. Sens. Lett. pp. 740–744 (2016)
42. Sommer, L.W., Schuchert, T., Beyerer, J.: Fast deep vehicle detection in aerial images. In: WACV. pp. 311–319 (2017)
43. Szeliski, R.: Computer Vision: Algorithms and Applications. Berlin, Heidelberg, 1st edn. (2010)
44. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: CVPR. pp. 10778–10787 (2020)
45. Tijtgat, N., Volckaert, B., Turck, F.D.: Real-time hazard symbol detection and localization using UAV imagery. In: VTC. pp. 1–5 (2017)
46. Wen, X., Shao, L., et al: Efficient feature selection and classification for vehicle detection. IEEE Trans. Circuits Syst. Video Technol. pp. 508–517 (2015)
47. Wu, Z., Suresh, K., et al: Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach. In: ICCV. pp. 1201–1210 (2019)
48. Xie, H., Chen, Y., Shin, H.: Context-aware pedestrian detection especially for small-sized instances with deconvolution integrated faster RCNN (DIF R-CNN). Appl. Intell. pp. 1200–1211 (2019)
49. Xie, H., Shin, H.: Two-stream small-scale pedestrian detection network with feature aggregation for drone-view videos. Multidimens. Syst. Signal Process. pp. 897–913 (2021)
50. Ye, J., Fu, C., et al: Darklighter: Light up the darkness for UAV tracking. In: IROS. pp. 3079–3085 (2021)
51. Yu, W., Yang, T., Chen, C.: Towards resolving the challenge of long-tail distribution in UAV images for object detection. In: WACV. pp. 3257–3266 (2021)
52. Zhang, C., Ge, S., et al: Accurate UAV tracking with distance-injected overlap maximization. In: MM. pp. 565–573 (2020)
53. Zhang, X., Izquierdo, E., Chandramouli, K.: Dense and small object detection in UAV vision based on cascade network. In: ICCV. pp. 118–126 (2019)
54. Zhu, P., Wen, L., et al: Vision meets drones: A challenge. CoRR p. abs/1804.07437 (2018)