2023

UIT-ADrone: A Novel Drone Dataset for Traffic Anomaly Detection

Tung Minh Tran
Vietnam National University, tungtm.ncs@grad.uit.edu.vn

Tu N. Vu
Vietnam National University

Tam Nguyen
University of Dayton, tnguyen1@udayton.edu

Khang Nguyen
Vietnam National University

UIT-ADrone: A Novel Drone Dataset for Traffic Anomaly Detection

Tung Minh Tran, Tu N. Vu, Tam V. Nguyen, *Senior Member, IEEE*, and Khang Nguyen

***Abstract*—Anomaly detection plays an increasingly important role in video surveillance and is one of the issues that have attracted various communities, such as computer vision, machine learning, and data mining in recent years. Moreover, drones equipped with cameras have quickly been deployed to a wide range of applications, starting from border security applications to street monitoring systems. However, there is a notable lack of adequate drone-based datasets available to detect unusual events in the urban traffic environment, especially in roundabouts, due to the density of interaction between road users and vehicles. To promote the development of anomalous event detection with drones in the complex traffic environment, we construct a novel large-scale drone dataset to detect anomalies involving realistic roundabouts in Vietnam, covering a large variety of anomalous events. Traffic at a total of three different roundabouts in Ho Chi Minh City was recorded with a camera-equipped drone. The resulting dataset contains 51 videos with total data traffic of nearly 6.5 h, captured across 206K frames with ten abnormal event types. Based on this dataset, we comprehensively evaluate the current state-of-the-art algorithms and what anomaly detection can do in drone-based video surveillance. This study presents a detailed description of the proposed UIT-ADrone dataset, along with information regarding data distribution, protocols for evaluation, baseline experimental results on our dataset, and other benchmark datasets, discussions, and paves the way for future work.**

***Index Terms*—Benchmark, convolutional neural networks (CNNs), drone-based surveillance, object detection, traffic anomaly detection.**

I. Introduction

UNUSUAL event detection is an active research topic in the fields of image processing and computer vision, which has attracted considerable attention from both academia and industry due to its many applications in real life. It is noteworthy that the detection of abnormal events, such as traffic collisions, violations, traffic accidents, fights, and crimes, has been one of the most crucial research topics of smart city transportation management systems in recent years.

An anomaly/outlier can be identified as activities or events that differ from what is expected, common, or normal [1], [2]. It means that it deviates significantly or has a low probability of occurring from some concept of normal, such as animals running into the roadway, truck accidents, stalled vehicles in transportation systems, defective products in the manufacturing industry, and the presence of a tumor in medicine. Thus, there is no fixed definition of anomalous actions or events in all the domains because the abnormality definition changes according to various application contexts, i.e., time, place, and scenarios. For instance, a person running at a park is usually a normal behavior but an abnormal event in other locations, such as a mall. Likewise, vehicles stopped near traffic lights are normal events when the traffic signal is red; however, they may be considered anomalous if the traffic light is green.

In recent years, unusual event detection has been a crucial component of the intelligent city transportation management system, primarily focusing on solving minority, unpredictable/uncertain, and rare events. It should be noted that detecting traffic anomalies involves multiple kinds of violations of regulations, such as driving in the wrong direction, illegal parking, car/motorcycle accidents, cyclist running on a pedestrian sidewalk, walking, fighting, robbing, and acts of vandalism. This is still a challenging problem due to the complex traffic environment, lightning conditions, dynamic weather conditions, lack of high-quality data, and the complexity of the traffic scene. Owing to recent considerable technological advancements, there is rapid growth in video surveillance networks that provide safety and security in public and private places, such as airports, streets, subway stations, hospitals, colleges, shopping malls, banks, companies, government buildings, and private homes. Moreover, drones are increasingly employed in various domains related to agriculture, the construction industry, border security, traffic monitoring system, and disaster area investigation.

When it comes to traffic surveillance systems, developing a the large dataset has become a challenge because abnormal objects are usually small interobject occlusions, and their visual features are not easily distinguishable, especially anomaly datasets from drone-based video surveillance to enhance public safety under real-world conditions. The main reason is that anomalous events occur infrequently and rarely due to the dependence on the changing visual context and the difficulty, unexpected cost, and laboriousness of sample collection in real-life situations. As a result, real-world anomaly datasets are severely

Manuscript received 7 March 2023; revised 4 May 2023; accepted 7 June 2023. Date of publication 20 June 2023; date of current version 28 June 2023. This work was supported by the Vietnam National University HoChiMinh City (VNU-HCM) under Grant B2023-26-01. *(Corresponding author: Khang Nguyen.)*

Tung Minh Tran, Tu N. Vu, and Khang Nguyen are with the University of Information Technology, Ho Chi Minh City 70000, Vietnam, and also with Vietnam National University, Ho Chi Minh City 70000, Vietnam (e-mail: tungtm.ncs@grad.uit.edu.vn; tamnguyen@udayton.edu; khangnttm@uit.edu.vn).

Tam V. Nguyen is with the Department of Computer Science, University of Dayton, College Park, Dayton, OH 45469 USA (e-mail: tamnguyen@udayton.edu).

The dataset is available online for noncommercial research at https://uit-together.github.io/datasets/UIT-ADrone/

Digital Object Identifier 10.1109/JSTARS.2023.3285905

imbalanced since the number of anomalies is much smaller than that of normal data. This problem leads to insufficiently labeled anomalous data, including suspicious human activities via training.

Motivated by the aforementioned challenges and the increasing demand for public safety and security, we particularly introduce a novel benchmark dataset captured by an aerial drone focusing on anomaly detection that is relevant to road traffic situations. Moreover, a comparative study with the existing state-of-the-art methods is also conducted in order to provide a challenging benchmark for real-time object detection and anomaly detection in aerial videos. Furthermore, to overcome the shortage of labeled anomaly data and to design a method that can extract meaningful features to effectively represent information from surveillance videos in a reasonable time frame, we propose a combination of deep transfer learning with unsupervised fine-tuning for anomaly detection from drone-based surveillance sequences. Transfer learning is based on a model trained with large-scale datasets from action recognition; that is, the convolutional neural networks (CNNs) trained on a particular dataset may be fine-tuned for a new dataset even if the scope is different without the need for relearning or providing new datasets. For example, a model that learned to identify trucks in a video stream can detect unseen cars without relearning the process.

The rest of this article is organized as follows. We first deliver related work on anomaly detection and review previous anomaly datasets captured by an aerial drone in Section II. We then detail the novel outlier detection relationships of our dataset in Section III. After that, experimental settings and several analytical experiments, including quality comparisons, are reported in Section IV. We then discuss the quantitative results and address the challenges associated with drone-view images in Section V. Finally, Section VI concludes this article.

II. Related Work

Detecting anomalous traffic events is a challenging problem in computer vision as it involves object detection, object tracking, and motion detection. To address the problem where anomalies only occupy a small amount of collected data, contrasting to normal data patterns that account for an overwhelming proportion of the data in various real-world applications, we review some methodologies that are closely related to deep learning drone-based video surveillance as well as existing drone-based datasets for object detection and anomaly detection tasks.

A. Methodologies for Anomaly Detection in Traffic Surveillance Videos

Concerning anomaly analysis in the traffic monitoring system, the main approach to anomaly problems using deep models consists of unsupervised and weakly supervised learning methods due to the limitations in the availability of annotated anomalous instances. Note that unsupervised learning techniques are used to detect abnormal events in surveillance videos, in which only normal data are available in the training step because anomalous events are diverse and difficult to capture. As a result, the number of anomalous samples collected for these methods needs to be increased compared with the normal objects. On the contrary, weakly supervised video anomaly detection methods that concentrate on training with numerous normal samples and a small number of category labels of unusual data patterns significantly improve learning accuracy compared to full unsupervised approaches. In particular, a comprehensive review of [3] focused solely on traffic anomalies. This survey presented different types of modern deep learning techniques applied in video clips to understand traffic violations and abnormalities in road traffic scenarios involving vehicles, pedestrians, and their interactions with the environment. Furthermore, this study reviewed computer-vision-based methods, frameworks, their applicability, implementation details, and limitations, discussed challenges, compared various benchmark datasets, identified gaps, and suggested future research directions. Next, in [4], U-Net used a generator to predict the next frame to detect anomalies in surveillance videos. After that, an optical flow constraint was proposed for the objective function constraining in terms of appearance and motion to ensure motion consistency for regular events in the training set to boost anomaly detection performance.

Moreover, it also used adversarial training to discriminate whether the prediction was actual or fake. In another one, Chang et al. [5] proposed a framework that dissociated spatial information and motion information using a two-stream architecture for video anomaly detection. In addition, the model utilized both reconstruction and prediction as auxiliary tasks for spatial and motion streams. This framework contained three key components. First, the first frame of the input video clips was fed into the spatial autoencoder network to detect anomalous objects with spatial features (e.g., scene and appearance). The given individual frame was encoded to a mid-level appearance representation by using the spatial encoder. Second, a motion autoencoder generated an RGB difference by inputting consecutive video frames, which could learn the temporal regularity. Moreover, its captured feature representation contained essential motion information. Third, variance-based attention in temporal autoencoder automatically assigned the importance weight to the moving part of video clips.

Concerning anomaly detection in aerial traffic surveillance, not much work has been done previously on unusual event detection for drone-based surveillance sequences. To be more detailed, a hybrid approach of [11] proposed to integrate space-time trajectories and semantic information of objects to build high-level knowledge for extracting complicated critical activities and events from drone-based video surveillance. In another work, Yang et al. [12] first classified safety-related abnormalities into three groups: 1) vehicles commit dangerous or illegal lane-changing behaviors; 2) vehicles slow down or stop unexpectedly or abruptly; and 3) vehicles blocked by vehicles in the crossing directions. Then, a functional approach was proposed to model temporal relations of time-to-collision safety indicators to detect safety-related anomalies from drone video surveillance. Next, an approach was made in [13] and proposed an architecture based on deep learning for contextual anomaly detection called CADNet. This method worked based on a variational autoencoder with a context subnetwork by exploiting

contextual information related to the environment from aerial video surveillance to find point anomalies and contextual anomalies. Then, Hamdi [18] proposed an unsupervised learning method based on deep end-to-end architecture for the detection of anomalies from drone-based surveillance. This method used only normal samples for the training phase, and the optical flow representations of abnormal samples were generated from consecutive original images in the testing phase. Most recently, Jin et al. [14] have introduced a method of anomaly detection in aerial videos using transformers by an encoder–decoder architecture called anomaly detection with transformer (ANDT). This framework aimed to treat adjacent frames as a sequence of triplets and then implemented a Transformer encoder to learn a spatiotemporal feature from the video sequence. Afterward, a decoder was applied to combine with the encoder to predict the next frame. Furthermore, ANDT focused solely on normal data in the training phase and identified an unknown or unpredictable event as an anomaly in the test phase.

B. Object Detection Models

We conduct a review of the literature related to real-time object detectors. As we can observe, real-time object detection plays an essential role in the field of computer vision to detect people, cars, bicycles, boats, and other objects in various contexts, such as traffic surveillance, robotics, and medical image analysis. In recent years, the top deep-learning-based object detection frameworks have been divided into two main categories, including one-stage architectures (e.g., YOLO [6], YOLOF [21], YOLOR [22], YOLOX [23], and YOLOv7 [25]) and two-stage architectures (e.g., Faster R-CNN [7]).

With regard to the one-stage-based methods, the **YOLO** [6] series is a representative one-stage network to enhance high accuracy and real-time speed. After that, this method continued to inspire further researches by making subsequent versions, and the detection performance improved steadily. Notably, the YOLO series has attracted considerable attention in the field of computer vision and various researchers in recent years. In particular, the architecture of **YOLOF** [21] consists of three main components: backbone, encoder, and decoder. YOLOF extracts only feature C5 level features from the backbone, without using extra features at other levels. Moreover, YOLOF replaces the RPN with a dilated encoder instead of using the feature pyramid network to extract features. In this architecture, the dilated encoder is designed to enlarge the receptive field. Furthermore, YOLOF addresses the imbalance problem in single-in-single-out encoders by applying a Uniform Matching strategy, which assigns each ground-truth box to the k nearest anchor points and k prediction boxes that are closest to it. This ensures that positive anchors are selected based on their proximity to the ground-truth boxes. In addition, uniform matching ensures that all the ground-truth boxes can be matched to the same number of positive anchors uniformly, regardless of their sizes. Next, **YOLOR** [22] is a useful unified model for multitasking purposes, especially real-time object detection. This network encoded implicit knowledge and explicit knowledge together and enabled the learned model to generate a unified representation to serve multiple tasks. Moreover, the kernel space in this framework was applied, which could be translated, rotated, and scaled to align each output kernel space of neural networks, prediction refinement, and multitask learning into the implicit knowledge learning process, and verified their effectiveness. Then, **YOLOX** [23] adopts the architecture of YOLOv3 [24] with DarkNet53 backbone by replacing the YOLO detect head with decoupled one to improve the convergence speed greatly. Moreover, the anchor-free and advanced labels were also assigned to this framework to improve object detector performance. In another one, **TPH-YOLOv5** [30], another variant of YOLOv5 [26], aims to detect multiscale objects by incorporating an additional prediction head. The model leverages transformer prediction heads and the convolutional block attention model to locate attention regions in dense object scenarios. Furthermore, various strategies, such as data augmentation, multiscale testing, multimodel integration, and leveraging extra classifiers, are employed in this framework to enhance the model's performance. Recently, **YOLOv7** [25] has been the trainable bag-of-freebies method to enhance the accuracy of real-time object detection. This method modified a more efficient ELAN module based on the YOLOv5 [26] algorithm and proposed a framework for auxiliary head training to enhance feature extraction, which improved accuracy and high performance.

Regarding the two-stage-based methods, **Faster R-CNN** [7] is a CNN-based method improved from the R-CNN architecture. The significant contribution of this model is the inference time at an approximate real-time speed. Faster R-CNN uses a pretrained CNN model to generate a feature map and bypasses the traditional region proposal algorithm of selective search. This feature map is then fed to the region proposal network (RPN) to identify the area recommendations and create predefined boxes called anchor boxes. The RPN is an alternative recommendation network to the selective search method. The anchor boxes are further classified and regressed. In addition, the nonmaximum suppression algorithm selects the overlapping anchors to ensure that the proposals do not contain overlapping boxes.

Concerning extensive experiments for object detection tasks, Nguyen et al. [32] investigated the impact of different deep learning object detection methods, including Faster R-CNN [7], RFCN [33], SNIPER [34], SSD [35], YOLOv3 [24], RetinaNet [36], and CenterNet [37] for object detection tasks in aerial images from drones. It has been demonstrated that the YOLO method is both feasible and effective based on experimental results, and the YOLO series is considered the optimal choice for real-time object detection applications. Next, Nguyen et al. [38] proposed an efficient approach (YALA) for learning to detect unseen (missing) objects. In this framework, a dual level of deep networks was designed to efficiently detect difficult objects in images by adopting Faster R-CNN [7] to train in the detection model and then training another Faster R-CNN model to tackle the unseen and challenging objects. Moreover, this pipeline leverages deep-learning-based multiscale detection for better performance. Further improvement of YALA was proposed in the study [39], named YADA, to improve object detection performance in images. This framework consisted of two stages: 1) data preparation during pretraining and 2) data residual posttraining. More specifically, lucid data synthesizing

TABLE I
COMPARISON OF PUBLICLY AVAILABLE DRONE-BASED DATASETS FOR ANOMALY DETECTION

Dataset attribute		Mini Drone [17]	UTT Drone [18]	Brutal Running [19]	AU-AIR [20]	Drone-Anomaly [14]	UIT-ADrone (Ours)
Dataset Size	No. of videos	38	–	–	1	59	51
	No. of frames	23,295	14,021	1,000	32,823	87,488	206,194 frames
	Length	16 – 24s per video	–	–	2 h	–	6.48 hours
Situations		parking lot	campus	–	roads	highways, roundabouts, etc.	roundabouts
Categories	People	✓	✓	✓	✗	✗	✓
	Vehicles	✗	✗	✗	✓	✓	✓
	Animals	✗	✗	✗	✗	✓	✗
No. of anomalous events		3	3	1	4	10	10
Resolution		224×224	227×227	227×227	1920×1080	640×640	1920×1080
Altitude		low altitude	low altitude	–	$10m$–$30m$	–	$50m$ - $70m$
Year		2015	2021	2021	2021	2022	2023

was applied in the data preparation to generate data by exploiting hard examples and embedding them in the same contextual locations. Furthermore, a dual-level deep network leveraged with these generated data was used by modifying Faster R-CNN [7] to train in a detection model. After that, another Faster R-CNN model was trained to detect hard objects in images.

C. Related Benchmark Drone-Based Datasets

This section introduces the public benchmark anomaly datasets developed by researchers in drone-based video surveillance, dealing with the complex traffic environment, as seen in Table I. Some of the mentioned datasets from aerial images focusing primarily on road traffic surveillance have been studied and published in recent years [15], [16]. However, in anomaly event detection, there are some drone datasets that exist [14], [17], [18], [19], [20]. To better understand different drone datasets, we briefly summarize them as well as are publicly available for research and useful for the comparison of different methods as follows.

1) Drone-Based Datasets for Vehicle Detection: The *VisDrone dataset* [29] consists of 10 209 images taken from drones in various locations at different heights for object detection task. There are ten predefined categories of interest, including pedestrian, people, bicycle, car, van, truck, tricycle, awning tricycle, bus, and motor. Moreover, some rare special vehicles are classified as "ignored regions" and "others," but they are not used in the evaluation. Out of 10 209 images, 6471 images are divided into training, 548 for validation, 1610 images for test-dev set, and 1580 images for testing. Furthermore, more than 540K bounding boxes of targets are annotated with ten predefined classes.

The *MONET dataset* [31] contains 53K images with 162K annotated bounding boxes captured with drones both day and night in a rural area near the city of Nicosia, Cyprus. Furthermore, there are three types of targets: vehicle, person, and ignore. Moreover, the dataset has many annotations that can also be used for multiple object-tracking problems.

The *UAVDT dataset* [9] contains 100 videos with a total duration of more than 10 h using drone cameras for object detection, single object tracking, and multiple object-tracking problems. The video sequences were recorded at 30 frames/s, with 1080×540 pixel resolution. Moreover, it was recorded at various locations in urban areas, including squares, arterial streets, toll stations, highways, crossings, and T-junctions. In addition, the videos cover different lighting conditions due to the time of the day and the weather conditions. In addition, approximately 80 000 frames in this dataset are labeled for more than 2700 vehicles with 0.84 million bounding boxes.

The *Stanford Drone dataset* [15] comprises about 10 000 trajectories with 929 499 samples in total. It was recorded from a drone's perspective with a length of 9 h at eight unique locations on the Stanford Campus. This dataset analyzes human trajectories in crowded scenes, such as pedestrians, bicycles, cars, skateboards, carts, and buses. Moreover, it has a high percentage rate of labeled pedestrians and cyclists, and their trajectory in time and space is also identified. At the same time, only approximately 7% of the labeled targets are cars.

The *DroneSURF dataset* [16] consists of 200 videos of 58 subjects captured with a drone camera for the problem of face recognition. The dataset includes a total of 411 451 samples. In addition, video footage has some challenges, such as motion, variations in pose, illumination change, background, altitude, and resolution. In addition, more than 786 000 face annotations are also provided for performance evaluation.

2) Drone-Based Datasets for Anomaly Detection: The *Mini-Drone dataset* [17] consists of 38 videos recorded with a Phantom 2 drone flying at low altitude in a parking lot for privacy protection. These videos are high resolution, with a duration ranging from 16 to 24 s each. The videos in this dataset are divided into three situation categories: normal, suspicious, and abnormal. Noticeably, these types are almost all identified by the actions of the persons involved in videos. More specifically, the normal actions in these videos relate to several events, such as people walking, getting in their cars, or parking correctly. The unusual activities include people fighting, a person falling down, and stealing. The suspicious cases represent situations where people behave suspiciously, which could distract the surveillance staff. For example, a person loitering in a parking lot can be considered as looking for a car/motorcycle to steal. Furthermore, the dataset comprises 15 training video sequences (9497 frames) and 23 testing video sequences (13 798 frames). The dataset is challenging because of changes in illumination, environmental variations, and different altitudes between videos.

In addition, the ground-truth annotations are provided for each video in the form of bounding boxes for each person and vehicle in each frame, which helps evaluate the performance.

The *UTT Drone dataset* [18] was captured with Mavic Air 2 from the DJI series with a total of 14 021 video frames (8933 for training and 5088 for testing). It contains seven folders for the training and 12 folders for the test. Particularly, the train folders contain only normal events, such as people walking on the lawn, whereas the test folders consist of both the normal and abnormal events. Unusual activities include running, fighting, and falling. However, the number of videos and the full video duration were not mentioned in detail in the original article.

The *Brutal Running dataset* [19] consists of 1000 samples in total captured with a Phantom 4 pro drone. There are 340 training samples and 660 samples for testing. The normal event consists of a girl walking outside, whereas the anomalous event occurs while running. Nevertheless, the number of videos, situations, and video length were not specifically mentioned in the original article.

The *AU-AIR-Anomaly dataset* [20] contains eight aerial videos of more than 2 h for traffic surveillance. Moreover, these videos were primarily captured at Skejby Nordlandsvej and P.O. Pedersens Vej roads (Aarhus, Denmark). Noticeably, this dataset was originally created for object detection tasks. Based on the dataset, Bozcan and Kayacan [13] annotated various abnormal events to detect anomalies in aerial videos. In addition, there are a total of 32 823 video frames covering four anomalous events, namely, a car on a bike road, a parked van in front of a building, a person on the road, and a bicycle on the road. Furthermore, frame-level ground truth is provided to evaluate the performance of state-of-the-art anomaly detection methods.

The *Drone-Anomaly dataset* [14] has a total of 59 untrimmed videos that are captured at seven different scenes in real-world environments, including highways, crossroads, bike roundabouts, vehicle roundabouts, railway inspection, solar panel inspection, and farmland inspection. Notably, aerial videos in this dataset were collected from YouTube and Pexels. In addition, the dataset comprises 37 training video sequences and 22 testing sequences with various real-world anomalous events. In addition, there are 87 488 video frames (51 635 for training and 35 853 for testing) in total, with each frame of 640×640 resolution and a frame rate of 30 frames/s. Moreover, the ground-truth annotations are provided for each testing video in the form of each anomalous event in each frame, which helps evaluate the performance. Nonetheless, the number of anomalies and the length of videos were not detailed in the original article.

III. Dataset Description

To tackle the limited availability of drone-based datasets with real anomalies for traffic anomaly detection, we construct a drone-view anomaly detection dataset, named UIT-ADrone. The drone took the video clips we chose to capture at realistic roundabouts, which are the most common place to capture all the road users present in a scene. Furthermore, the interaction between traffic participants is particularly high at these intersections, especially motorcycles. Our dataset has a wide range of challenges. In more detail, the object scale in aerial images varies dramatically due to the substantial change in flight altitude and distance between the drone and objects of interest. In addition, drone-based video surveillance objects (e.g., cars, bicycles, and motorcycles) have illumination changes, occlusion of independently moving objects, and complex backgrounds. In addition, anomalies mostly occur for a short span of time in video drones.

A. Data Acquisition

We have conducted videos captured by the aerial perspective with a camera to collect practical traffic data for detecting traffic abnormalities in Ho Chi Minh City, Vietnam. The used drone is a DJI MAVIC MINI 2, recording at 30 frames/s, with a resolution of 1920×1080 pixels ranging from 50 to 70 m in height at different times of the day. More specifically, the videos of the drone are recorded at two roundabouts on the campus of the Vietnam National University, Ho Chi Minh City, and at one public roundabout in Ho Chi Minh City. Note that the public roundabout has an especially high traffic volume, with a variety of motorbikes on the streets. Moreover, Fig. 1 shows a scene at the roundabout in our dataset.

B. Dataset Statistics

The UIT-ADrone dataset consists of 51 videos with a total of 206 194 extracted video frames covering various anomalous events. The entire video is approximately 6.50 h long, recorded in complex real-world scenarios, and they pose significant new challenges, such as complex scenes, high density, occlusion of moving objects, lighting conditions, small objects, and large camera motion. Furthermore, it contains ten abnormal events related to various types of violations of regulations, including crossing the road at the wrong lane, walking under the street, driving in the wrong roundabout, illegally driving on the sidewalk, illegal left turn/turn right, illegally parking in the street, carrying bulky goods, parking on the sidewalk, driving in the opposite directions, and falling off motorcycles.

In Table I, we compare the UIT-ADrone dataset with current public datasets for anomaly detection problems, which are similar to the UIT-ADrone dataset: Mini-Drone [15] and UTT Drone [18]. It is essential to notice that the datasets in this table only provide three types of abnormal events, involving some pedestrians and cars. Therefore, there is not much interaction between road users and vehicles (e.g., cars, buses, trucks, and bikes), as well as interactions between objects of interest, especially motorcycles. In contrast to the existing datasets, the UIT-ADrone dataset consists of ten different types of anomalous events and a more representative distribution of the types of road users in public urban interchanges. Moreover, the videos captured with a drone have a duration of nearly 6.50 h of data. Furthermore, the detailed descriptions of anomalous event types in our dataset are given in Table II. Notably, a frame can have more than one outlier event because of the huge diversity of situations that are encountered on real-world roundabouts, especially motorbikes. Thus, our dataset is much more suitable for detecting abnormal events from drone-based traffic videos.

Fig. 1. Visualization of traffic scenes included in the UIT-ADrone (Ours) dataset.

TABLE II
STATISTICS OF THE UIT-ADRONE (OURS) DATASET

Types of abnormal events	Number of actions
Crossing the road at the wrong line	80
Walking under the street	344
Driving in the wrong roundabout	686
Driving on the sidewalk	145
Illegal left turn/turn right	28
Illegal parking in the street	225
Carrying bulky goods	144
Parking on the sidewalk	68
Driving in the opposite directions	214
Falling off motorcycles	1

C. Annotation

To generate ground-truth data for the purpose of evaluating different models to detect traffic anomalies, we use a tool called Supervisely assign frame-level labels. This tool is browser based and supports advanced functions, such as drawing a bounding box or tracking the objects of interest in a video drone. Since the frames we label are sequential at a frame rate of 30 frames/s and the position of objects of interest changes little from frame to frame, Supervisely is an easy-to-use tool for annotating challenge data. Furthermore, some screenshots of the Supervisely tool are illustrated in Figs. 2 and 3. This tool is available at https://supervise.ly/. In addition, the UIT-ADrone dataset comprises bounding boxes for the training and testing sets for object detection, following the format of the MS COCO dataset [27], which is standard for object detection. Therefore, our dataset contains train.json and test.json for the training and testing sets, respectively. For anomaly detection, testing labels are organized as arrays (.npy). It means that the video is equal to the array. The indexes for frames start at 0 and end at a total of frames subtraction of 1. Each element in the array is set to 0 as normal or 1 as abnormal. Notably, only normal data are available in the training step for these unsupervised tasks in our study. In addition, we also carry out cross-checking between annotators to check for error labels based on the consensus of annotators.

IV. PROTOCOLS AND BASELINE RESULTS

In this section, we demonstrate experimental results to verify the challenges and effectiveness of our dataset based on a set of state-of-the-art algorithms. Concretely, protocols and baseline results have been provided for the task of object detection and traffic anomaly detection on our dataset and two public datasets. The performance of the methods on standard datasets for both of these tasks is illustrated in Tables III and IV. Furthermore, we also conduct extensive cross-dataset experiments to investigate the cross-dataset adaptability of our dataset. The experimental outcomes are presented in Tables V and VI.

A. Protocols

The proposed UIT-ADrone dataset is downsampled and divided into the number of abnormal and normal snippets for training and testing sets. In particular, the training set only includes normal snippets, whereas the testing set consists of normal and abnormal snippets. The resulting dataset has a total of 1497 snippets (592 for training and 905 for testing), or the equivalent of 206 194 video frames. In more detail, there are 59 186 frames for the training set and 147 005 frames for the testing set. Moreover, there are 63 485 ground-truth annotations provided for testing snippets in the form of bounding boxes

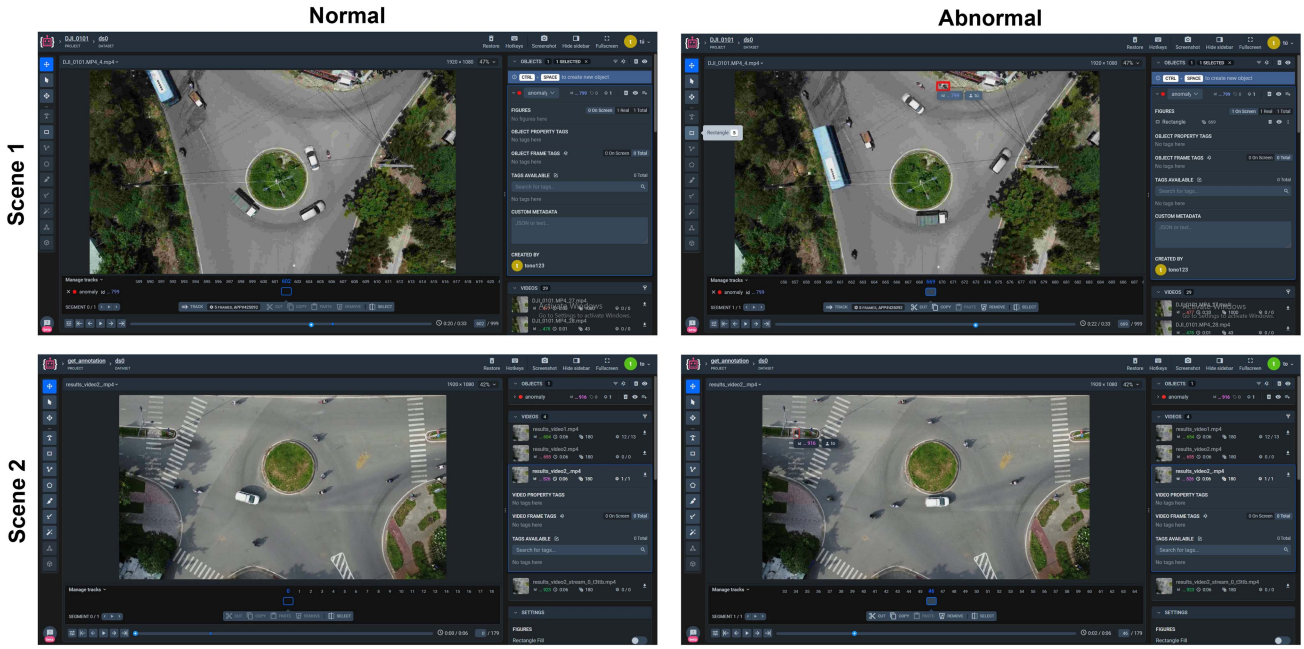

Fig. 2. Example images from two traffic scenes of the UIT-ADrone dataset by using Supervisely. The left column shows normal frames, and the right column demonstrates abnormal frames. Note that, red boxes denote anomalies in abnormal frames. Please zoom by 400% in the electronic version.

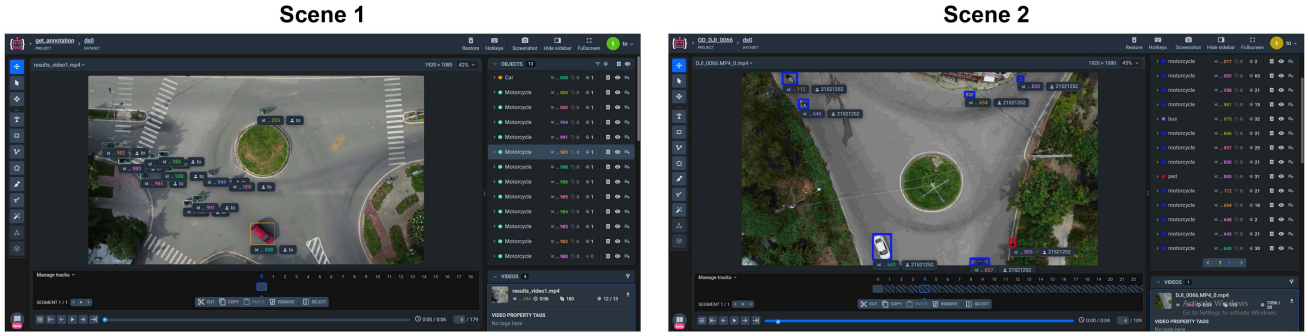

Fig. 3. Image samples for object detection from two traffic scenes of the UIT-ADrone dataset by using the Supervisely annotation tool. Please zoom by 400% in the electronic version.

around each abnormal event in each extracted video frame. Noticeably, each anomalous event (object) is also labeled with a tracking number. Furthermore, a single frame can have more than one labeled anomaly.

B. Experimental Results

In this section, we discuss the experimental results of our UIT-ADrone dataset with two common tasks: object detection and anomaly detection. Based on related works, thoroughly empirical studies with five object detectors consisting of Faster R-CNN [7], YOLOF [21], YOLOR [22], YOLOX [23], and YOLOv7 [24] are performed on the UAVDT dataset [9] and our dataset for the object detection task. Moreover, we also conduct an evaluation of three current state-of-the-art anomaly detection algorithms based on deep architectures, namely, Future Frame Prediction [4], ANDTs [14], and Spatiotemporal Dissociation [5] on two anomaly datasets, including the Drone-Anomaly [14] dataset and our dataset, which follow the same setup as other similar unsupervised video anomaly detection studies.

1) Metrics: We use a frame-based receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC) [8] to evaluate the performance of experimental methods for anomaly detection problems. In addition, mAP [10] is a very important metric used to measure object detection performance in our benchmark. mAP metric will be calculated by using official MS-COCO api. It should be noted that the higher the value of AUC, as well as the value of mAP, is, the better the model's performances will have. Furthermore, Fig. 4 presents the results of our experiments on the Drone-Anomaly [14] and our datasets in terms of the ROC-AUC metric at the frame level. Moreover, we provide the snapshots of some correct cases and failure cases on our dataset from our experiments, as seen in Fig. 5.

TABLE III
COMPARISON OF EXPERIMENTAL METHODS IN TERMS OF AVERAGE PRECISION VALUES ON THE UIT-ADRONE (OURS) AND THE UAVDT DATASET (%).

Method	AP_{bus}	AP_{truck}	AP_{car}	$AP_{tricycle}$	AP_{van}	AP_{bike}	AP_{ped}	$AP_{motorbike}$	AP_{50}	AP_{75}	mAP
Faster R-CNN [7]	30.90	26.30	23.00	16.00	6.80	1.40	5.00	6.40	31.90	11.30	**14.50**
	7.80	1.90	27.90	–	–	–	–	–	22.30	12.90	**12.50**[1]
YOLOF [21]	31.80	26.00	24.30	18.30	9.80	1.20	2.80	5.30	32.80	11.60	**14.90**
	1.70	1.20	26.50	–	–	–	–	–	20.90	7.80	**9.80**
YOLOR [22]	32.60	28.80	26.90	21.20	11.20	1.19	5.40	8.31	35.50	13.70	**16.90**
	11.30	2.65	42.90	–	–	–	–	–	31.00	21.10	**19.00**
YOLOX [23]	30.00	21.00	19.30	13.60	3.50	1.30	2.60	5.30	26.50	9.40	**12.10**
	5.30	0.00	21.30	–	–	–	–	–	20.00	6.10	**8.90**
YOLOv7 [24]	28.30	29.20	26.20	17.80	8.28	0.27	3.50	6.14	34.40	11.50	**15.00**
	23.00	3.14	34.40	–	–	–	–	–	37.00	20.02	**20.10**

[1] The result was cited from [28]. The results of state-of-the-art methods on the UAVDT dataset are denoted by gray cells.

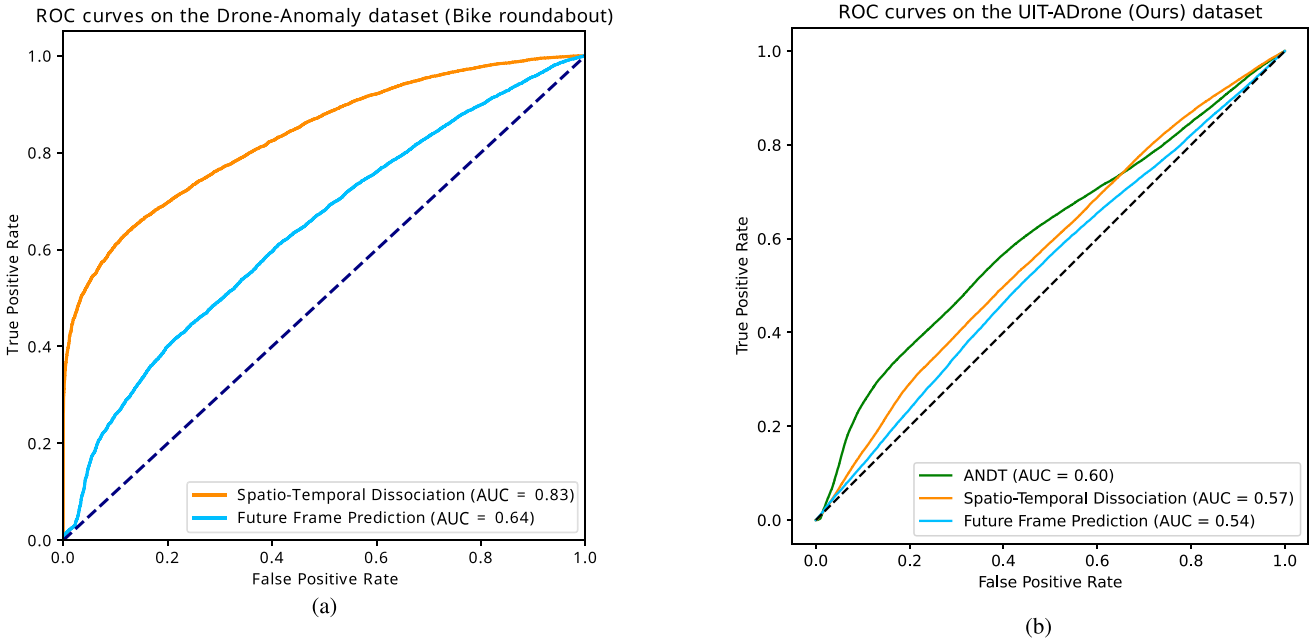

Fig. 4. Experimental results of Future Frame Prediction [4], ANDT [14], and Spatiotemporal Dissociation [5] methods based on ROC-AUC metric at the frame level of the Drone-Anomaly [14] and the UIT-ADrone (Ours) datasets. (a) Visualization plot of ROC-AUC score on the Drone-Anomaly dataset [14]. (b) Visualization plot of ROC-AUC score on the UIT-ADrone (Ours) dataset.

TABLE IV
PERFORMANCE COMPARISON OF ANOMALY DETECTION METHODS IN TERMS OF AVERAGE AUC METRIC AT FRAME LEVEL ON THE DRONE-ANOMALY [14] AND OUR DATASETS (%)

Method	AUC	
	UIT-ADrone (Ours)	**Drone-Anomaly [14]**
Future Frame Prediction [4]	53.56	64.00
ANDT [14]	60.50	82.20[1]
Spatio-Temporal Dissociation [5]	57.05	79.50

[1] The result was quoted from [14] for the roundabout scene.

TABLE V
RESULTS OF CROSS-DATASET ADAPTATION EXPERIMENTS IN TERMS OF AVERAGE AUC METRIC AT FRAME LEVEL ON THE DRONE-ANOMALY AND THE UIT-ADRONE (OURS) DATASETS WITH FINE-TUNING AND WITHOUT FINE-TUNING (%)

Source → target	AUC
w/ Fine-tuning	
UIT-ADrone → Drone-Anomaly [14]	83.40
Drone-Anomaly [14] → UIT-ADrone	55.30
w/o Fine-tuning	
UIT-ADrone → Drone-Anomaly [14]	76.86
Drone-Anomaly [14] → UIT-ADrone	52.84

TABLE VI
COMPARISON OF THE NUMBER OF SAMPLES BETWEEN DRONE-ANOMALY [14] AND UIT-ADRONE (OUR) DATASETS

Dataset	Number of samples
Drone-Anomaly (the roundabout scene) [14]	26,377
UIT-ADrone (Ours)	**206,194**

The emphasize number of frames of the UIT-ADrone dataset which is much more than that of the Drone-Anomaly dataset at the same scene (roundabout scene).

2) Object Detection: Table III uses mAP metric to show a performance comparison of various algorithms, including the Faster R-CNN [7], YOLOF [21], YOLOR [22], YOLOX [23], and YOLOv7 [24] models for object detection task on the UIT-ADrone (Ours) and UAVDT [9] datasets. In general, out of five well-known methods, the mAP of the YOLOR [22] on our dataset and that of the YOLOv7 [24] method on the UAVDT dataset [9] are the highest at nearly 17.00% and at just over 20.00%, respectively. In contrast, the figure of the YOLOX [23]

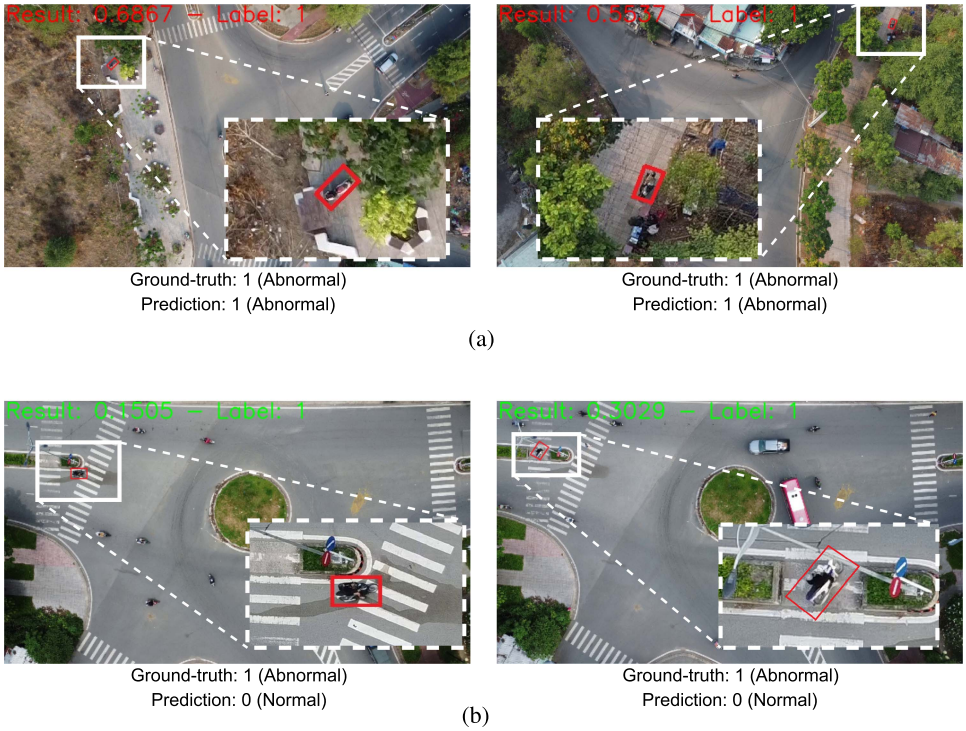

Fig. 5. Correct and failure cases for anomaly prediction on the UIT-ADrone (Ours) datasets. (a) Correct cases for frame prediction on the UIT-ADrone dataset. (b) Failure cases for frame prediction on the UIT-ADrone dataset.

method is the lowest on the UIT-ADrone and UAVDT datasets at nearly 9.00% and at roundly 12.00%, respectively. Moreover, the experiments also show that one-stage YOLO family detectors perform well in contexts of aerial images rather than two-stage detectors, such as Faster R-CNN [7]. As regards the UIT-ADrone dataset, the figures of the YOLOR [22] model surpass the other state-of-the-art methods in most of the classes on the UIT-ADrone dataset, ranging from about 1.00% to nearly 4.00%. However, this method achieves the lowest result on AP_{bike} class at nearly 2.00%. Furthermore, the mAP of the YOLOv7 [24] method is slightly higher than the Faster R-CNN [7] model at nearly 15.00% and 14.50%. Regarding the UAVDT dataset [9], it is clear that the largest percentage of the YOLOv7 [24] method accounts for 20.10%, and this figure is slightly higher than that of the YOLOX [23] method at 19.00%. Moreover, compared with the rest of the models, the performance result of the YOLOR [22] method is better across all the classes, ranging from about 2.60% to over 20.00%. On the other hand, the percentage of the YOLOX [23] and YOLOF [21] methods is the lowest at approximately 9.00% and at roundly 10.00%, respectively. Notably, the object detection result of the YOLOX [23] method on AP_{truck} class is 0.00%. From the results presented in this table, it can be seen that the YOLOR [23] and YOLOv7 [24] models obtain the best object detection performance for both single and multiple object detection compared with the existing methods on two benchmarks. Interestingly, the performance of the YOLOv7 [24] method on the UAVDT dataset [9] is higher than that of the YOLOR [23] method; however, its performance is less than that of the YOLOR method on our UIT-ADrone dataset. Theoretically, YOLOR [23] is the model that is proposed to train multitasks for gaining implicit knowledge to serve other tasks. In the case of the UAVDT dataset [9], having images captured at an angle of under 90° with the horizontal, and the objects are much smaller than our UIT-ADrone dataset; implicit knowledge from pretrained models may not work well, leading to a worse performance than the YOLOv7 [24] model. The images' context of our UIT-ADrone dataset seems to be similar to knowledge trained on the YOLOR [23] method; therefore, the implicit knowledge works well. Conversely, the YOLOv7 [24] model aims to generalize on different contexts of datasets without prior knowledge; thus, its performance may be worse than that of the YOLOR [23] method. In addition, experimental results obtained on two benchmark datasets demonstrate the effectiveness of the YOLOv7 [24] method in the case of less training data and faster computation. In contrast, the YOLOR [23] algorithm requires large data to train. Therefore, this method significantly improves the result of common classes on two datasets. As regards the UIT-ADrone dataset, besides the common classes, there are some classes with small objects along with complicated calculations; then, the YOLOR [23] model gives better results. Notably, the number of object classes detected in our dataset is almost three times more than that of object classes in the UAVDT dataset [9] (eight classes compared with three classes). Still, the YOLOv7 [24] model on the UAVDT dataset [9] has a much higher mAP value (about 5.00%) than the achieved result of this method on the UIT-ADrone dataset at 20.10% and 15.00%. The detailed figures' analysis demonstrates that our anomaly dataset is challenging for the current state-of-the-art algorithms for object detection tasks in drone-based video surveillance systems.

3) Anomaly Detection: Table IV shows performances of three prominent methods, namely, Future Frame Prediction [4], ANDT [14], and Spatiotemporal Dissociation [5] on two abnormal drone datasets, namely, Drone-Anomaly [14] and UIT-ADrone based on AUC metric at the frame level. Overall, there are considerable differences in results between the state-of-the-art methods on these datasets. The ANDT method's result is the highest figure of the three prominent methods on two datasets at over 82.00% for the Drone-Anomaly dataset and at 60.50% for the UIT-ADrone dataset. Furthermore, the ANDT method is higher than that of the other ones ranging from about 3.00% to over 10.00% on two anomalous datasets. Similar to the ANDT method, the next most substantial percentage of the Spatiotemporal Dissociation method is at 79.50% for the Drone-Anomaly dataset and at 64.00% for the UIT-ADrone dataset. By contrast, the figure of the Future Frame Prediction method is the lowest on these datasets at 64.00% for the Drone-Anomaly dataset and 53.00% for the UIT-ADrone dataset. Based on the analysis of these experimental results, the results of these well-known methods on the UIT-ADrone dataset are much lower than the results of the Drone-Anomaly dataset. It is clear that our dataset is very challenging due to the complex traffic environment and diversity of anomalous data. Obviously, the issue of anomalous event detection in drone-based video surveillance is still challenging, depending on the realistic conditions and environment.

4) Transfer Learning: Our extensive experiments aim to fine-tune as well as without fine-tuning for detecting anomalies in frames on cross-dataset adaptation. The detailed analysis of the performance and its comparison in terms of AUC metric at frame level with the state-of-the-art method, namely, Spatiotemporal Dissociation [5] on two sources, and destination datasets, including the Drone-Anomaly [14] and UIT-ADrone datasets, are also reported. To be more detailed, we first train the model on training images from the source dataset and then perform a transfer learning setting by loading the pretrained weights learned from the training image in the source dataset to continue to learn the training image patterns from the target dataset. After that, the final weights will be evaluated on the testing set of the target dataset. By refining the learned representations for anomaly detection, the Drone-Anomaly and UIT-ADrone datasets serve as source and target datasets, respectively. Similarly, we also experiment without fine-tuning (inference dataset B directly from the model trained on dataset A, and *vice versa*) for detecting anomalies in frames on these datasets.

Table V shows the performance evaluation of the Spatiotemporal Dissociation [5] method for anomaly detection in the two sources and destination datasets, namely, the Drone-Anomaly [14] and UIT-ADrone datasets. It is noteworthy that the experimental method with fine-tuning has superior results to those without fine-tuning on cross-dataset adaptation. Specifically, the performance of the setting UIT-ADrone → Drone-Anomaly [14] with the Spatiotemporal Dissociation [5] method on the Drone-Anomaly's testing set with fine-tuning surpasses the performance of the setting without fine-tuning by over 6.50%, at 83.40% compared to at 76.86%. Likewise, the performance of the setting Drone-Anomaly [14]→UIT-ADrone witnesses the same conclusion with transfer learning experiments: the setting of fine-tuning achieves the higher AUC score (about 2.50%) than that of without fine-tuning at 55.30% and 52.84%. Moreover, from the results presented in this table, we see that the performance of the setting UIT-ADrone → Drone-Anomaly [14] with the Spatiotemporal Dissociation [5] method on the Drone-Anomaly's testing set outperforms the performance of the setting without loading the learned weights increased by nearly 4.00%, at 83.40% compared to at 79.50% of the previous experimental result from Table IV. The learned models can explain this outcome through the load weights learned from the UIT-ADrone dataset, which provides significant prior knowledge because our dataset includes numerous images captured from roundabout scenes. Although humble AUC scores obtained from the setting without fine-tuning compared to those with fine-tuning, the results are also pretty good and even acceptable in the context of no training. This, therefore, leads to better performance in small-scale datasets, such as Drone-Anomaly (captured at the roundabout scene similar to our dataset context). On the contrary, the performance of the setting Drone-Anomaly [14]→UIT-ADrone with the Spatiotemporal Dissociation [5] baseline does not improve on the testing set of our dataset at 55.30%, nearly 2.00% lower than the previous experimental result from Table IV of 57.00%. We assume that the Drone-Anomaly dataset has a much smaller number of samples than that of our dataset (26 377 samples compared to 206 194 samples, according to Table VI) as well as the lack of variety of unusual event types in the same context, resulting in poor generalization of the trained model due to the rapid growth of video surveillance data, especially data captured with drone in the complex traffic environment. It is essential to note that loading learned weights from the Drone-Anomaly dataset generates some noise due to insufficient training data. Thus, the trained model has difficulty continuing to learn patterns from large-scale datasets, such as the UIT-ADrone dataset.

V. Discussion

In this article, we create an annotated aerial video dataset consisting of 51 video sequences involving three realistic traffic scenes. Our UIT-ADrone dataset expands the scope of anomaly detection research in real-world applications by covering a large variety of anomalous events with characteristics in Vietnam. In addition, we extensively validate the existing methods in order to provide a benchmark for this task. In addition, owing to the complexity and diversity of real-world scenarios, the most significant challenge in traffic anomaly detection problems is that available data are highly imbalanced toward normality (i.e., nonanomalous), leading to the fact that the availability of anomalous cases may be limited and may evolve over time due to external factors. Within such scenarios, it is impossible to take into account all the unseen abnormal examples, so our dataset is very challenging for well-established state-of-the-art methods to detect unusual events.

According to the above experimental findings, we find that compared with natural images, new challenges appear in aerial images based on real scenes, including dense distribution, miniature objects with only a few pixels, large aspect ratios, arbitrary

orientations, and camera motion. These characteristics make deep learning models, such as CNN-based and transformer-based, face challenges for aerial object detection as well as detect anomalous events in drone-based video surveillance, especially the recorded surveillance footage of dense populations in urban environments. Furthermore, small objects are easily fooled by complex background noise interference, thus, in turn, increasing the difficulty of accurate object detection and detection of unusual events. In addition, anomalies in surveillance footage are difficult to anticipate, as the prior knowledge about these anomalies is usually limited or even unavailable. In addition, when new categories of anomalies emerge, the data acquisition environments are extremely diverse, and the labeling of training data for novel categories is complex and prohibitively expensive. Furthermore, abnormal event detection is more than just dependent on circumstances and context but also depends on the appearance of the objects and their movements in real scenes. Moreover, challenges in anomaly detection with massive volumes of aerial videos include inconsistent behavior of different types of anomalies, camera movements, variable spatial resolution due to changes in flight altitude, and handling imbalanced distribution of normal and abnormal data, in which normal events often account for an overwhelming proportion. Therefore, state-of-the-art methods are typically trained only on normal data while being tested on both the normal and abnormal data, which are difficult to adapt to various monitoring scenarios. Furthermore, the analysis of the extensive experiments shows the superiority of the proposal for cross-domain adaptivity on our dataset, in which many anomalous events with temporal dynamics exist. It is clear that these experimental outcomes demonstrate that our dataset can apply deep-learning-based transfer learning for drone-based anomaly detection. Moreover, since the main focus of this article is to introduce a novel dataset for anomaly detection in drone images, we have not yet evaluated the accuracy of various kinds of anomalous events.

VI. Conclusion

In this article, we presented our efforts to build a novel dataset for the real-time detection of anomalous events in aerial traffic surveillance at three various locations, in which the primary context is roundabout. We contributed the large-scale dataset of aerial videos, named UIT-ADrone, with specific applicability to detect anomalous events in the complicated background and various object sizes. The proposed dataset contained 51 original videos of ten abnormal events recorded on various roundabouts and at different times of the day. Video frames were captured across over 206 000 frames, with 63 485 anomalous frames annotated. Furthermore, extensive empirical results performed on various publicly available benchmarks demonstrated that it is challenging to track small objects. The experiments showed that it is feasible to detect anomaly frames in real-life applications.

Future work will consider increasing the number of environmental contexts to increase the diversity of anomalous event types and more experiments that can be performed by evaluating the accuracy of different abnormal events, as well as tracking anomalous events from the UIT-ADrone dataset. Finally, by sharing our dataset, we hope that researchers will push the limitations of the existing methods for object detection as well as outlier event detection in aerial videos.

References

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, pp. 1–58, 2009.

[2] T. M. Tran, T. N. Vu, N. D. Vo, T. V. Nguyen, and K. Nguyen, "Anomaly analysis in images and videos: A comprehensive review," *ACM Comput. Surv.*, vol. 55, no. 7, pp. 1–37, 2022.

[3] K. K. Santhosh, D. P. Dogra, and P. P. Roy, "Anomaly detection in road traffic using visual surveillance: A survey," *ACM Comput. Surv.*, vol. 53, pp. 1–26, 2020.

[4] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6536–6545.

[5] Y. Chang et al., "Video anomaly detection with spatio-temporal dissociation," *Pattern Recognit.*, vol. 122, 2022, Art. no. 108213.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[7] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[8] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[9] D. Du et al., "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 370–386.

[10] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, pp. 98–136, 2015.

[11] D. Cavaliere, V. Loia, A. Saggese, S. Senatore, and M. Vento, "A human-like description of scene events for a proper UAV-based video content analysis," *Knowl.-Based Syst.*, vol. 178, pp. 163–175, 2019.

[12] D. Yang, K. Ozbay, K. Xie, H. Yang, F. Zuo, and D. Sha, "Proactive safety monitoring: A functional approach to detect safety-related anomalies using unmanned aerial vehicle video data," *Transp. Res. C: Emerg. Technol.*, vol. 127, 2021, Art. no. 103130.

[13] I. Bozcan and E. Kayacan, "Context-dependent anomaly detection for low altitude traffic surveillance," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2021, pp. 224–230.

[14] P. Jin, L. Mou, G.-S. Xia, and X. X. Zhu, "Anomaly detection in aerial videos with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5628213.

[15] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 549–565.

[16] I. Kalra, M. Singh, S. Nagpal, R. Singh, M. Vatsa, and P. B. Sujit, "DroneSURF: Benchmark dataset for drone-based face recognition," in *Proc. IEEE 14th Int. Conf. Autom. Face Gesture Recognit.*, 2019, pp. 1–7.

[17] M. Bonetto, P. Korshunov, G. Ramponi, and T. Ebrahimi, "Privacy in mini-drone based video surveillance," in *Proc. IEEE 11th Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2015, vol. 4, pp. 1–6.

[18] S. Hamdi, "Deep learning anomaly detection for drone-based surveillance," doctoral dissertation, Univ. Technol. Troyes, Troyes, France, 2021.

[19] S. Hamdi, S. Bouindour, H. Snoussi, T. Wang, and M. Abid, "End-to-end deep one-class learning for anomaly detection in UAV video stream," *J. Imag.*, vol. 7, 2021, Art. no. 90.

[20] I. Bozcan and E. Kayacan, "AU-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 8504–8510.

[21] Q. Chen, Y. Wang, Y. Tong, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13039–13048.

[22] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "You only learn one representation: Unified network for multiple tasks," 2021, *arXiv:2105.04206*.

[23] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[24] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[25] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7464–7475.
[26] J. Yao, J. Qi, J. Zhang, H. Shao, J. Yang, and X. Li, "A real-time detection algorithm for Kiwifruit defects based on YOLOv5," *Electronics*, vol. 10, 2021, Art. no. 1711.
[27] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
[28] W. Yu, T. Yang, and C. Chen, "Towards resolving the challenge of long-tail distribution in UAV images for object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3258–3267.
[29] D. Du et al., "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 213–226.
[30] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2778–2788.
[31] L. Riz et al., "The MONET dataset: Multimodal drone thermal dataset recorded in rural scenarios," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2545–2553.
[32] K. Nguyen, N. T. Huynh, P. C. Nguyen, K. D. Nguyen, N. D. Vo, and T. V. Nguyen, "Detecting objects from space: An evaluation of deep-learning modern approaches," *Electronics*, vol. 9, 2020, Art. no. 583.
[33] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 379–387.
[34] B. Singh, M. Najibi, and L. S. Davis, "SNIPER: Efficient multi-scale training," in *Proc.Int. Conf. Neural Inf. Process. Syst..*, 2018, pp. 9333–9343.
[35] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis.*2016, pp. 21–37.
[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[37] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6569–6578.
[38] K.-D. Nguyen, K. Nguyen, D.-D. Le, D. A. Duong, and T. V. Nguyen, "You always look again: Learning to detect the unseen objects," *J. Vis. Commun. Image Represent.*, vol. 60, pp. 206–216, 2019.
[39] K.-D. Nguyen, K. Nguyen, D.-D. Le, D. A. Duong, and T. V. Nguyen, "YADA: You always dream again for better object detection," *Multimedia Tools Appl.*, vol. 78, pp. 28189–28208, 2019.

Tung Minh Tran received the B.Sc. degree in information system from the University of Science, Vietnam National University, Ho Chi Minh City (VNUHCM), Ho Chi Minh City, Vietnam, in 2001, and the M.Sc. degree in geographic information system from the University of Technology, VNUHCM, in 2005. He is currently working toward the Ph.D. degree in computer science with the University of Information Technology, VNUHCM.

In 2007, he joined the University of Finance and Marketing, Ho Chi Minh City. His research interests include data science, computer vision, and artificial intelligence.

Tu N. Vu received the B.Sc. degree in computer science from the University of Science, Vietnam National University, Ho Chi Minh City (VNUHCM), Ho Chi Minh City, Vietnam, in 2022. He is currently working toward the M.Sc. degree in artificial intelligence with Chonnam National University, Gwangju, South Korea.

In 2017, he joined the University of Information Technology, VNUHCM. His research interests include computer vision and deep learning.

Tam V. Nguyen (Senior Member, IEEE) received the Ph.D. degree in information technology from the National University of Singapore, Singapore, in 2013.

He was a Research Scientist and a Principal Investigator with the ARTIC Research Centre, Singapore Polytechnic, Singapore. He was also an Adjunct Lecturer with the National University of Singapore. He is currently an Assistant Professor with the Department of Computer Science, University of Dayton, Dayton, OH, USA. His research interests include computer vision, applied deep learning, multimedia content analysis, and mixed reality.

Khang Nguyen received the B.Sc. and M.Sc. degrees in computer science and the Ph.D. degree in information technology from the University of Science, Vietnam National University, Ho Chi Minh City (VNUHCM), Ho Chi Minh City, Vietnam, in 1990, 1995, and 2012, respectively.

He is currently the Vice-President of the University of Information Technology, VNUHCM. His research interests include artificial intelligence and computer vision.