DroneChase: A Mobile and Automated Cross-Modality System for Continuous Drone Tracking

Neel R Vora[1], Yi Wu[2], Jian Liu[2], Phuc Nguyen[1]
[1]The University of Texas at Arlington, [2]The University of Tennessee, Knoxville
neelrajeshbhai.vora@uta.edu;ywu83@vols.utk.edu;jliu@utk.edu;vp.nguyen@uta.edu

ABSTRACT

This paper presents *DroneChase*, an automated sensing system that monitors acoustic and visual signals captured from a nearby flying drone to track its trajectory in both line-of-sight and non-line-of-sight conditions under mobility settings. Although drone monitoring has been an active research topic, most of the existing monitoring systems focus only on line-of-sight conditions and do not perform well under blockage conditions. Inspired by the human ability to localize objects in the environment using both visual and auditory signals, we develop a mobile system that integrates the information from multiple modalities into a reference scenario and performs real-time drone detection and trajectory monitoring. Our developed system, controlled by the Raspberry Pi platform, collects acoustic signals from 6 hexagonal channels placed 5 cm away from each other and video signals from an HD RGB camera. The monitoring system is placed in a moving vehicle and is able to track the drone even when it is flying/hovering behind the bush or trees. Furthermore, the portability of the system enables continuous chasing of the drone, allowing for uninterrupted monitoring and tracking even while on the move. In addition, the proposed system performs reliably in both day and night conditions.

CCS CONCEPTS

• **Computing methodologies** → **Tracking**.

KEYWORDS

drone localization, transfer learning, multimodal drone tracking

1 INTRODUCTION

Drones have gained immense popularity in recent years due to their versatility and ability to be used in various fields, including aerial photography, surveillance, delivery, and agriculture. Nevertheless, the proliferation of drones has led to concerns about their potential use in illicit activities such as espionage, smuggling, and even terrorist attacks [13]. Developing effective drone localization systems is therefore critical to ensure the safety and security of critical infrastructure, public spaces, and sensitive locations. In addition to providing early warning and detection of unauthorized drone activity, such systems can also assist in tracking and intercepting rogue drones and provide valuable intelligence on the activities of hostile actors. As a result, there is a growing need for robust and reliable drone localization systems that can operate in a range of environments.

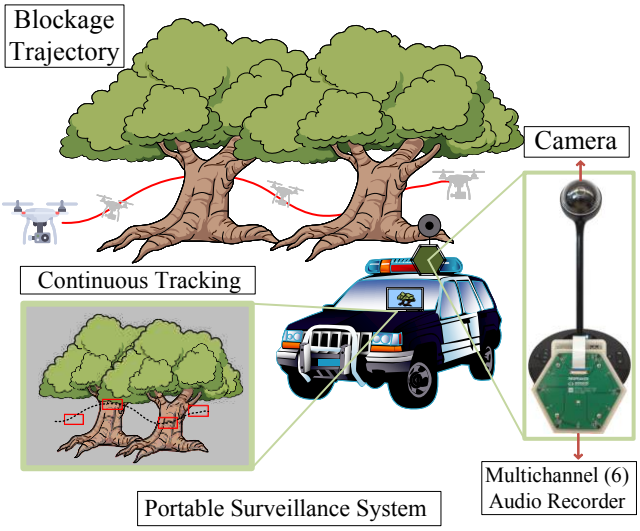

Figure 1: DroneChase's concept: drone localization in low-light, blockage, and mobility conditions [1].

Prior Research. Traditional vision-based approaches can localize the drone with high accuracy [2, 6]. However, they require the drone to be within the line of sight of the cameras and have good lighting conditions. This can limit the capabilities of the system, particularly when the drone is obscured by objects such as trees or buildings. Despite the use of multiple cameras at different angles, these approaches still struggle to localize the drone when it is behind objects. Similarly, radar-based approaches also require line-of-sight conditions and lose effectiveness under physical obstacles [22]. RF-based approaches are proposed [3, 11] as an alternative

DroNet '23 , June 18, 2023, Helsinki, Finland

ACM ISBN 979-8-4007-0210-5/23/06.
https://doi.org/10.1145/3597060.3597237

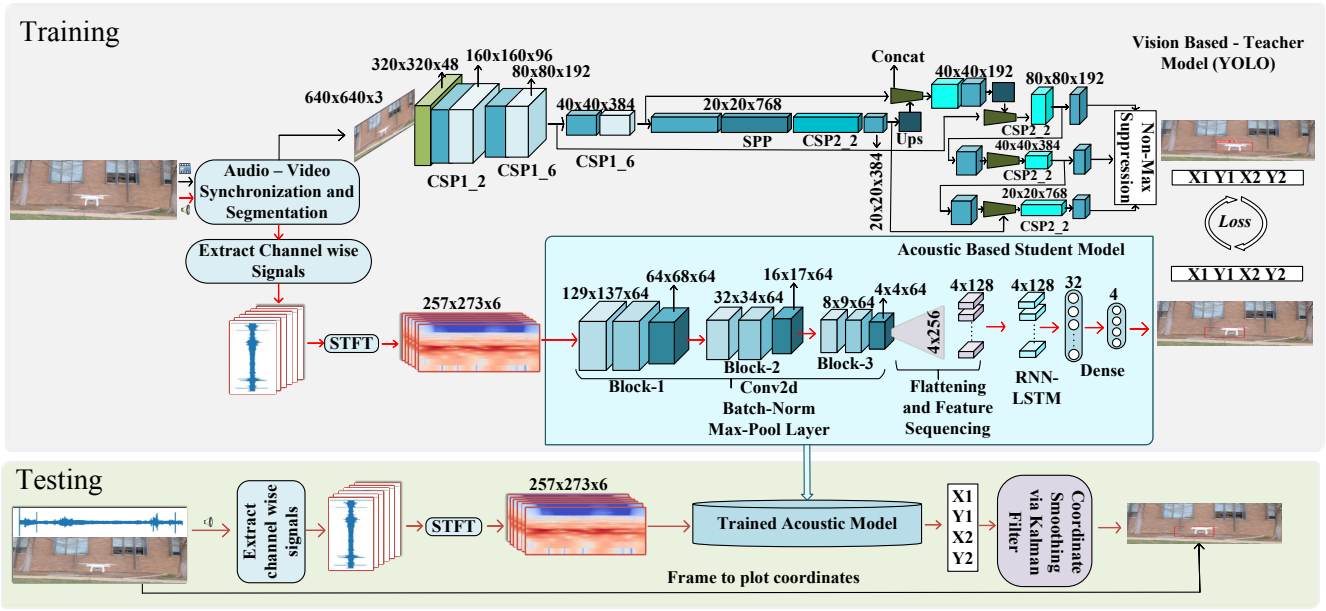

Figure 2: System Overview.

method to localize drones. However, these methods either necessitate the installation of wireless transceivers on the drone, which involves re-engineering the flight controller, or the need to know the communication frequency/signal generated by the drone that can be utilized for localization, but it is often not readily available as companies tend to conceal this information. Additionally, several acoustic-based approaches have been proposed. [5, 9, 19].

Nevertheless, these methods suffer from certain limitations that hinder their practical application. Specifically, they rely on cumbersome signal recording systems that lack mobility or require modifications to the drone itself, such as adding speakers to generate sound pulses for localization purposes. Therefore, there is a need for alternative localization techniques that can overcome these limitations and allow the drone to operate effectively in a broader range of environments with increased coverage. Rather than relying on mathematical properties of sound that constrain the design of the recording device, we opted for a data-driven approach that harnesses the capabilities of machine learning.

In this paper, we propose a data-driven mobile system, named *DroneChase*, as illustrated in Fig. 1 to detect and localize the drone with high precision under blockage and mobility conditions. The proposed method involves creating a student model that emulates a higher-dimensional teacher model, thereby facilitating continuous drone localization. The objective is to develop a system capable of learning auditory-visual correspondences in a self-supervised manner, enabling classical object detection tasks such as drawing bounding boxes around the target, using only acoustic information from a drone. In this paper, we made the following contributions:

- The proposed approach avoids the need for manual labeling. Instead, we employed a self-supervised training method that automates the labeling process for the localization of drones.
- We developed a machine-learning model to localize the drone and evaluated our model with real-world settings. The system was placed on top of a car and tested in various outdoor environments, including low light and blockage scenarios.
- We developed affordable, portable hardware using Raspberry Pi and Seeed ReSpeaker to capture multi-channel surround sound data.

2 SYSTEM OVERVIEW

Our proposed system, *DroneChase*, offers a portable solution for drone tracking that can be mounted on a vehicle, allowing for effective chasing of the drone. Its portability ensures that the system can provide extensive coverage, making it highly practical to real-world implementation

In lieu of the time-consuming and manual collection of ground truth data, we present a novel approach that employs a student-teacher-based architecture to exploit the labeling capacity of the vision model and train the acoustic model with those labels, thereby allowing for the continual tracking of a drone via only acoustic signals. Removing the need for manual collection of labels makes drone tracking systems scalable.

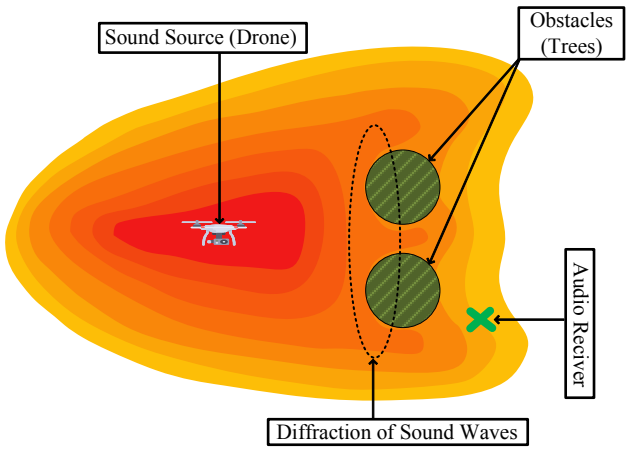

Figure 3: Sound waves propagation around obstacles

As illustrated in Fig 2, our system has two phases: the training phase and the testing phase. During the training phase, the system trains an acoustic model in a self-supervised manner using a vision model that localizes drones from video frames. In the testing phase, the acoustic model operates independently, continuously localizing drones without requiring any visual input. The system is further composed of three subsystems:

Training vision model: We are using a transfer learning approach to train the YOLOv5 [7] on about 10,000 annotated images of drones. The model is able to achieve high precision in localizing the drone in a certain environment. Later this trained model is used as a teacher model which will generate a ground truth guide for the acoustic model. For this, we first perform video resampling to ensure uniformity in frame rate and synchronization with the sound signal, each frame of the processed video is fed to the teacher model, and the generated coordinates are used for loss function during the training phase of the acoustic model.

Training acoustic model: The acoustic model takes spectrograms of sound signals as an input instead of the raw signal as the spectrogram provides information in both time and frequency domains and thus can gain more spatial information. We begin with resampling all six channels' audio signals to maintain uniformity of sample rate. Each channel is divided into segments corresponding to the frame rate of the video, i.e., each 1-sec audio is divided into 30 segments. Then we apply Short-Time Fourier Transform (STFT) on each segment of every channel to compute their spectrograms. The spectrograms of an audio segment of each channel are stacked together and fed to Multi-input Convolution Recurrent Neural Network (CRNN) to output a bounding box similar to the teacher model. CRNNs incorporate a feedback mechanism that allows them to better capture temporal dependencies and utilize this information to have a smoother tracking system

Testing acoustic model: During testing, the audio signals first pass through the same pre-processing procedures as in training. Following that spectrograms are generated and fed to the trained student model, The model will then generate the coordinates of the bounding box without any requirement for visual input. We then apply to coordinate smoothing via the Kalman filter to further stabilize the bounding-box movement.

3 APPROACH

Our approach is grounded on the physical principle of wave diffraction, which enables sound waves to bend and propagate beyond obstacles, as illustrated in Fig. 3. Additionally, sound waves also convey spatial information about their source. By leveraging these properties of sound waves, we have proposed a machine-learning-based approach that can localize the source of the sound, even when it is hidden behind obstacles thus allowing us to overcome the limitations posed by obstructed line-of-sight and has the potential to significantly enhance drone tracking capabilities in such conditions.

The learning problem for our system is modeled using a student-teacher framework, whereby the system is trained simultaneously using both video frames and multi-channel acoustic signals. The training procedure of the network across multiple sensing modalities, i.e., vision, and acoustic is described in detail in this section. This approach enables the acoustic student network to learn how to localize the drone guided by the visual teacher network in a self-supervised manner.

3.1 Vision Network

We used a transfer learning approach to train YOLOv5 [7] which was pre-trained on the Microsoft Common Objects in Context dataset. Its architecture is a convolutional neural network composed of a backbone and a detection head. The backbone is based on the CSP-Darknet53 architecture, which is a modified version of Darknet53 used in YOLOv3 [14]. CSPDarknet53 comprises 53 convolutional layers, including residual blocks that aid in training and performance enhancement. The backbone also contains downsampling layers that reduce feature map spatial resolution while increasing their depth. The overall architecture consists of 77 convolutional layers in both the backbone and detection head. During training, we froze the backbone layers and fine-tuned them on our annotated drone dataset.

After completing the training of YOLOv5, it is utilized as a teacher network to create a stream of pseudo-labels for the student network during its training. To enable this, the data is initially prepared by resampling the

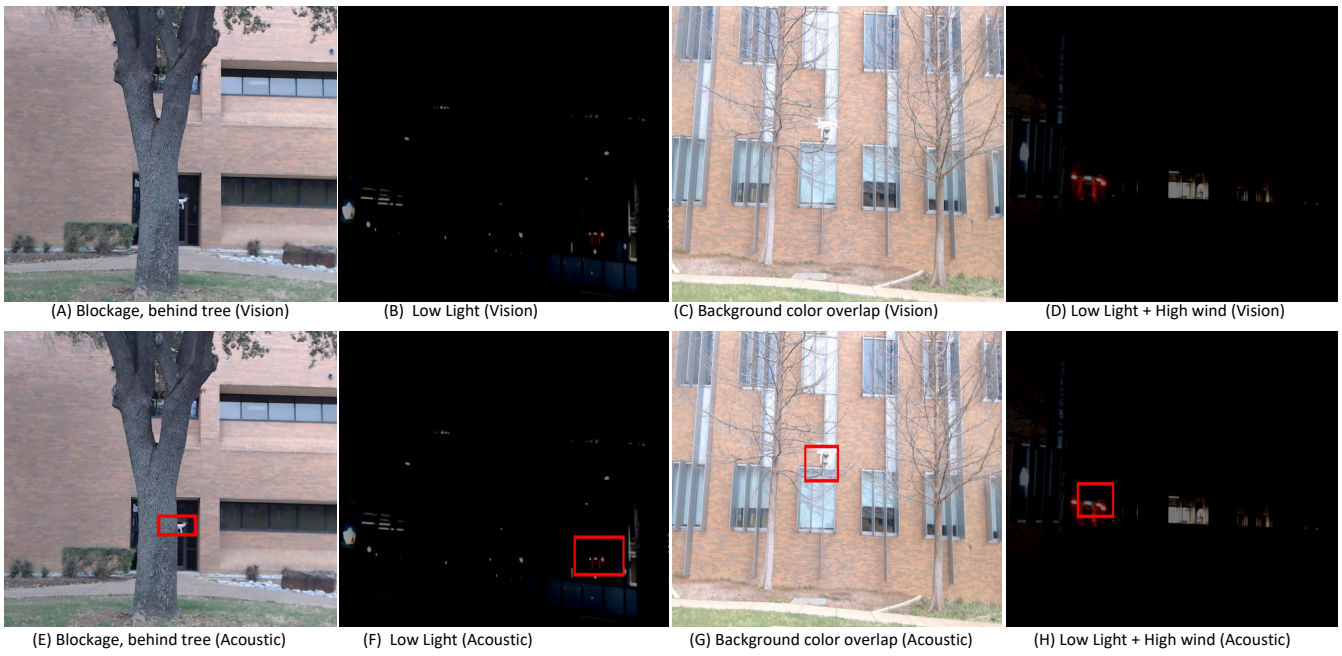

Figure 4: Performance comparison of Our System with YOLO.

video to 30 frames per second (fps), ensuring uniform synchronization with audio segments. This step is essential in making certain that video recorded from any kind of hardware can be used with the system.

The processed video is then decomposed into several $T = 1s$ video clips, which are further divided into 30 frames. These frames are continuously fed into the trained teacher model, which, in turn, generated output labels comprising coordinates of the bounding box. These pseudo-labels are then sent to the student network, which is simultaneously training on audio data.

3.2 Acoustic Network

We approached object detection from acoustic by framing it as a regression task. Acoustic data are collected from six different channels of our hardware as shown in Fig. 5 thus providing more spatial information. We first begin by resampling 6-channel audio data to 48000KHz, Each of these channels is divided into $T = 1s$ chunks, and these are further divided into 30 segments that correspond to each frame from video data. Each segment is then converted into a spectrogram through STFT. Utilized the hamming window function with window size 512, and 4 Hops, this gives us a spectrogram of shape [257×273]. Spectrograms of a segment from all 6 channels are stacked together and fed to the student network.

The student network utilized in our proposed approach is a CRNN, which comprises three 2D convolutional layers. Each convolutional layer is followed by a batch normalization layer and a max pooling layer. The output generated from the last convolutional layer is flattened to create a feature sequence, which is then fed into a Long Short-Term Memory (LSTM) layer. The LSTM layer is then followed by two dense layers to produce the final output.

Loss Function. During the training of our learning model, the objective is to minimize the loss functions that are designed to represent the error between the predicted bounding box and the corresponding ground truth. In this approach, we utilized the Intersection Over Union (IoU) loss [23], which considers the spatial relationships between the predicted and ground truth bounding boxes. This loss function is relatively robust to variations in object size, position, and orientation, making it an effective choice in object detection tasks. IoU loss for bounding boxes is defined as:

$$L_{IoU}(y,t) = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{Area(y \cap t)}{Area(y \cup t)} \quad (1)$$

Here, y represents the predicted bounding box, t represents the target bounding box, N is the batch size, and $Area$ calculates the area of a bounding box.

Optimization. Furthermore, the network is trained to utilize the Adam optimizer algorithm as proposed by [8] with learning rate set to 0.0001. In order to avoid overfitting we used L1 & L2 Kernel regularizer.

3.3 Smoothing via Kalman Filter

The coordinates/size of the bounding boxes regressed from the network is inevitably jittery, which is caused by ambient noises in the environments as well as the instability of the network. We thus utilize Kalman Filter [20]

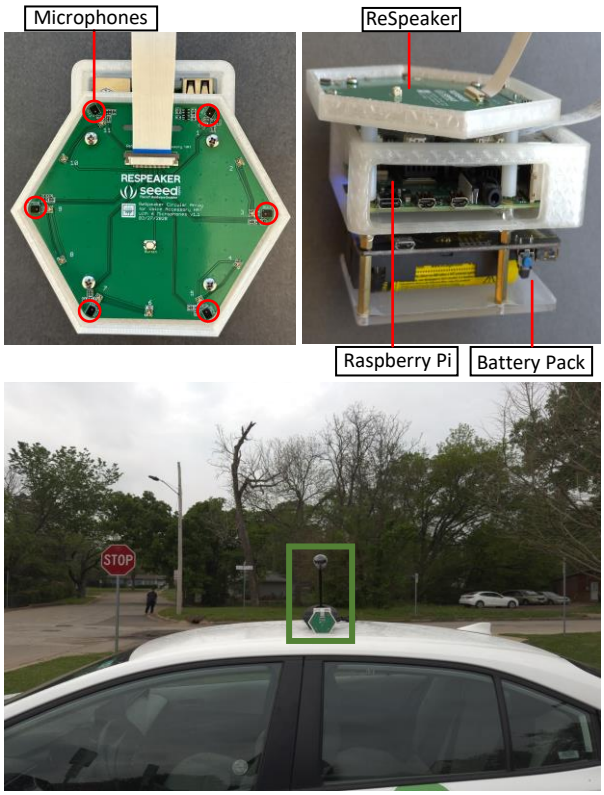

Figure 5: DroneChase's Prototype

to smooth the coordinates between adjacent frames. Given the vector that represents the bounding box at frame t, we define its state vector $s_t = [x^t, y^t, w^t, h^t]$, where x^t and y^t represent the coordinate, w^t and h^t stand for the width and height. A state-space model defining the bounding box movement thus can be represented as $s_t = As_{t-1}$, and the bounding box $z_t = Hs_t$. As the time gap between adjacent frames is relatively short, we define $A = 1$, $H = I$, and the process & measurement noise co-variances are empirically set to 0.01 & 0.05, respectively. The smoothed bounding box can then be derived as $\hat{s}_t$, where $\hat{s}_t$ is the optimal state estimate.

4 EVALUATION

4.1 System Implementation

We implemented an *DroneChase* system as shown in Fig.5 using Seeed Respeaker and Raspberry Pi to record audio signals. To make the hardware compact we used a 5V battery pack with raspberry pi. This device is capable of collecting 6 channels of surround sound at a maximum sample rate of 48KHz with an SNR of 59 dB using 6 MEMS microphones and consumes only 2.7W. For recording video, we are using Logitech Conference Cam BCC950, which records 1080p video at 30 fps.

In our study, we employed a total of six microphones to capture spatial information, in contrast to traditional mono or stereo recorders. Our findings indicate that this approach resulted in a significant improvement in the overall quality of the recorded audio, as evidenced by the data presented in Table 1. To support the circuit, Raspberry Pi, and battery, we developed a customized 3D-printed case, which provided both ergonomic and sturdy support for the recording system.

4.2 Experiment Methodology

In this study, a dataset of drone flying was collected consisting of 60 minutes of video and audio data in different environmental conditions. The drone flies in a 7m $\times$ 7m $\times$ 7m cubical area following a random trajectory to avoid any bias in the data. We used DJI Phantom-4 drone for our experiment purpose. The video was recorded at 30 fps, with 1920 $\times$ 1080 resolution which is widely used by most cameras, and audio was recorded at a sampling rate of 48KHz for higher fidelity.

Data Split. We split video frames and audio clips into three parts: 72K for training, 18K for validation, and 18K for testing. One audio clip is 33.3ms i.e., 1600 samples and each frame is of 1920 $\times$ 1080 resolution. This dataset was used to train and evaluate the proposed system.

4.3 Evaluation Results

Evaluation Metric. To evaluate our method, we used the traditional object detection evaluation metric, Average Precision (AP). We report the AP at IoU with 0.5 and 0.75 thresholds. The formula for AP_k is given by:

$$AP_k = \frac{1}{|G|} \sum_{g=1}^{|G|} [IoU(y,t) \geq t] \tag{2}$$

where G is the set of ground truth objects, $|G|$ is the cardinality of G, Iverson bracket notation $[IoU(y,t) \geq t]$ evaluates to 1 if the condition is met i.e., if $IoU \geq$ threshold t, and 0 otherwise.

We also calculated the average of AP across IoU thresholds from 0.3 to 0.75 with an interval of 0.05 as:

$$A_{ve}AP_{30:75} = \frac{1}{|K|} \sum_{i=1}^{|K|} AP_{K[i]} \tag{3}$$

where K is the set of IoU thresholds, and $|K|$ is the cardinality of K which is 10 in this case (IoU thresholds from 0.30 to 0.75 with an interval of 0.05), and $AP_{K[i]}$ is the Average Precision at IoU threshold $K[i]$ which is calculated using Eq(2).

Result Analysis. We evaluated our system with various channel configurations, as shown in Table 1. It turned out that using all 6 channels gives a more prominent result, mainly because it separates orientation better and has more spatial information. In our case, 2 channels (mic3,mic6), 4 channels (mic1,mic6,mic2,mic3), and

Approach	AP_{ave}	$AP_{0.5}$	$AP_{0.75}$
Mono	33.27	30.03	11.38
Stereo	37.25	37.66	16.87
3 Channels	41.32	40.93	16.26
4 Channels	49.065	48.808	21.59
5 Channels	53.40	50.73	22.98
6 Channels	**61.20**	**59.43**	**25.63**

Table 1: Compared Average Precision (AP) across multiple channels. Higher AP suggests better results.

6 channels show better improvement in performance with the best result on using all 6 channels.

We further evaluated our system on different settings that include Low light, Blockage, High winds and etc. As shown in Table 2, our standalone acoustic system performed significantly well compare to vision model YOLOv5. As illustrated in Fig.4, the corresponding result of the vision-based approach's fail cases.

Approach	AP_{ave} Blockage	AP_{ave} Normal
YOLOv5	8.43	87.16
Our System	42.95	61.20

Table 2: Compared Average Precision (AP) between YOLO and our system. Higher AP indicates better performance.

5 RELATED WORK

Vision-based Drone Localization. Vision-based drone localization systems detect drones from video frames captured by cameras leveraging various deep-learning-based object detectors e.g., [14, 15]. and [10] combine Faster R-CNN with ResNet-101 to detect drones from long-range surveillance videos, similarly [12] localize drone swarms using cameras mounted on a headset. Alternatively, spatial-temporal attention and U-net-based frameworks have also been proven effective on drone localization [2, 6]. However, vision-based methods require good line-of-sight and lighting conditions and may raise privacy concerns.

RF- and Radar-based Drone Localization. RF- and radar-based drone localization research utilizes RF signals transmitted between the drone and operator to determine location, with good performance but requires mounting additional wireless transceivers on the drone, and more importantly, the communication frequency of drones must be known while manufacturers tend to conceal this information. [3, 11]. Radar-based methods emit electromagnetic waves and analyze reflected signals, but similar to vision-based methods, they require line-of-sight and lose effectiveness with physical occlusions. [4, 21, 22].

Acoustic-based Drone Localization. There has been activating research on drone localization leveraging the Time Difference of Arrival (TDoA) of acoustic signals [5, 16–18]. However, these systems have significant limitations. For instance, [16] require a cumbersome hardware setup involving 40 microphones, while [18] proposes an indoor solution that can only be effective if the drone has limited movement space. Additionally, [5, 17] requires a large setup space with a minimum separation distance of 20 meters between microphones. Alternatively, [19] mount speakers that transmit acoustic pulses on the drone, while ground-based microphones detect the pulses and determine the drone's location. Despite the promising results, mounting additional hardware on the drone inevitability affect its mobility and requires considerable hardware changes. [9] proposed an audio-visual-based approach, but similarly, the inconvenient hardware setup (i.e., 30 cameras) largely limits its usage scenario. Compared to existing solutions, our system is low-cost, highly portable, and does not require any hardware modification to the drone, allowing for easy deployment in various scenarios and reducing the need for additional engineering efforts.

6 CONCLUSION AND FUTURE WORK

We present *DroneChase*, an automated cross-modality learning system for self-supervised drone localization. It uses a student-teacher model and YOLOv5-based automatic labeling. We utilized a cost-effective mobile audio sensor for data collection and developed a quantitative evaluation method to assess the system's performance under various environmental conditions. We conducted experiments in challenging scenarios such as high wind, night-time, blockages, and crowded campus settings, as well as a moving car to mimic a dynamic environment. Our findings show the system's capacity to improve tracking in challenging conditions, We provide a live demonstration of DroneChase's real-time outdoor drone tracking performance [1].

However, our current system is limited to tracking a single drone at a time. In future work, we plan to expand our approach to enable the tracking of multiple drones simultaneously. We also aim to collect data in even more severe environmental conditions to further demonstrate the robustness of our system. This study also paves the way for developing a generalized model that can track different objects that emit sound waves. We intend to investigate more sophisticated audio processing techniques and explore the use of other sensing modalities such as radar so that it can enhance our system's tracking capabilities.

Acknowledgments. This material is based partly upon work supported by the National Science Foundation under Award Numbers 2132112 and 2152357.

REFERENCES

[1] Dronechase's demo. https://youtu.be/p_WuN-3Xzlo, 2023.

[2] M. W. Ashraf, W. Sultani, et al. Dogfight: Detecting drones from drones videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7067–7076, 2021.

[3] S. Basak and B. Scheers. Passive radio system for real-time drone detection and doa estimation. In *2018 International Conference on Military Communications and Information Systems (ICMCIS)*, pages 1–6. IEEE, 2018.

[4] I. Bouzayene et al. Scan radar using an uniform rectangular array for drone detection with low rcs. In *2019 IEEE 19th Mediterranean Microwave Symposium (MMS)*, pages 1–4. IEEE, 2019.

[5] X. Chang, C. Yang, et al. A surveillance system for drone localization and tracking using acoustic arrays. In *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 573–577. IEEE, 2018.

[6] C. Craye and S. Ardjoune. Spatio-temporal semantic segmentation for drone detection. In *2019 16th IEEE International conference on advanced video and signal based surveillance (AVSS)*, pages 1–5. IEEE, 2019.

[7] G. Jocher et al. ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations, 2021.

[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[9] H. Liu, Z. Wei, et al. Drone detection based on an audio-assisted camera array. In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pages 402–406. IEEE, 2017.

[10] M. Nalamati, A. Kapoor, et al. Drone detection in long-range surveillance videos. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2019.

[11] P. Nguyen, T. Kim, et al. Towards rf-based localization of a drone and its controller. In *Proceedings of the 5th workshop on micro aerial vehicle networks, systems, and applications*, pages 21–26, 2019.

[12] M. Pavliv, F. Schiano, et al. Tracking and relative localization of drone swarms with a vision-based headset. *IEEE Robotics and Automation Letters*, 6(2):1455–1462, 2021.

[13] T. Pledger. The role of drones in future terrorist attacks. *Association of the United States army*, 26, 2021.

[14] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[15] S. Ren, K. He, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[16] A. Sedunov, D. Haddad, H. Salloum, A. Sutin, N. Sedunov, and A. Yakubovskiy. Stevens drone detection acoustic system and experiments in acoustics uav tracking. In *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*, pages 1–7. IEEE, 2019.

[17] Z. Shi, X. Chang, et al. An acoustic-based surveillance system for amateur drones detection and localization. *IEEE transactions on vehicular technology*, 69(3):2731–2739, 2020.

[18] Y. Sun, W. Wang, et al. Aim: Acoustic inertial measurement for indoor drone localization and tracking. 2022.

[19] W. Wang, L. Mottola, et al. Micnest: Long-range instant acoustic localization of drones in precise landing. 2022.

[20] G. Welch, G. Bishop, et al. An introduction to the kalman filter. 1995.

[21] J. Yang, X. Lu, et al. A cylindrical phased array radar system for uav detection. In *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pages 894–898. IEEE, 2021.

[22] J. Zhao, X. Fu, et al. Radar-assisted uav detection and identification based on 5g in the internet of things. *Wireless Communications and Mobile Computing*, 2019, 2019.

[23] D. Zhou, J. Fang, X. Song, C. Guan, J. Yin, Y. Dai, and R. Yang. Iou loss for 2d/3d object detection. In *2019 International Conference on 3D Vision (3DV)*, pages 85–94. IEEE, 2019.