Archangel: A Hybrid UAV-Based Human Detection Benchmark With Position and Pose Metadata

YI-TING SHEN1,2, (Student Member, IEEE), YAESOP LEE1,2, HEESUNG KWON3, DAMON M. CONOVER3, SHUVRA S. BHATTACHARYYA1,2, (Fellow, IEEE), NIKOLAS VALE3, JOSHUA D. GRAY4, G. JEREMY LEONG5, KENNETH EVENSEN6, AND FRANK SKIRLO6

1ECE Department, University of Maryland, College Park, MD 20742, USA
2UMIACS, University of Maryland, College Park, MD 20742, USA
3DEVCOM Army Research Laboratory (ARL), Adelphi, MD 20783, USA
4Fibertek Inc., Herndon, VA 20171, USA
5Department of Energy, Washington, DC 99352, USA
6Defense Threat Reduction Agency (DTRA), Fort Belvoir, VA 22060, USA

Corresponding author: Yi-Ting Shen (ytshen@umd.edu)

This work was supported in part by the Defense Threat Reduction Agency (DTRA), and in part by the U.S. Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory (ARL) under Grant W911NF2120076.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Learning to detect objects, such as humans, in imagery captured by an unmanned aerial vehicle (UAV) usually suffers from tremendous variations caused by the UAV’s position towards the objects. In addition, existing UAV-based benchmark datasets do not provide adequate dataset metadata, which is essential for precise model diagnosis and learning features invariant to those variations. In this paper, we introduce Archangel, the first UAV-based object detection dataset composed of real and synthetic subsets captured with similar imagining conditions and UAV position and object pose metadata. A series of experiments are carefully designed with a state-of-the-art object detector to demonstrate the benefits of leveraging the metadata during model evaluation. Moreover, several crucial insights involving both real and synthetic data during model optimization are presented. In the end, we discuss the advantages, limitations, and future directions regarding Archangel to highlight its distinct value for the broader machine learning community.

INDEX TERMS UAV-based object detection, human detection, UAV-based benchmark dataset, position metadata, synthetic data, model optimization.

I. INTRODUCTION

With the recent rapid advancement in edge computing technology coupled with resource-constrained mobile platforms, particularly unmanned aerial vehicles (UAVs) with electro-optical (EO) sensor payloads, a wide range of UAV-enabled applications have been more prevalent. Notable examples include UAV-enabled search and rescue in disaster management [1], aerial surveillance and reconnaissance for civilian and military purposes [2], precision agriculture [3], traffic analysis [4], and intelligent transportation applications [5]. Lately, owing to the remarkable progress of artificial intelligence and machine learning technology, tailored to distinct constraints of small UAV platforms, these UAV-based applications have been frequently providing promising solutions and successfully achieving the goals and operational requirements of their corresponding applications. Central to the above UAV-based applications are streamlined plug-ins that can effectively interrogate real-time imagery, captured with UAVs, and provide image/video analytics relevant to UAV-based scene understanding, particularly object detection and recognition.

Recently, extensive efforts in object detection and recognition have led to extraordinary advances in the perception
accuracy in various challenges associated with large-scale object detection benchmarks that were captured primarily with ground-based cameras [6], [7], [8]. Compared to ground-based object detection and recognition, UAV-based object detection poses unique and severe challenges, as UAV flight inevitably results in a wider range of variations in the conditions for capturing images, including the altitudes and viewing angles of cameras, system turbulence, and weather events. These variations lead to more drastic variations in object appearances/attributes, and thus pose additional challenges to onboard detection models in the location and recognition of objects of interest. In general, changes in object appearances/attributes, caused by varying the image collection conditions, entail three major dependencies: (1) pose dependency caused by changes in the UAV position or camera viewing angle, (2) scale dependency owing to the distance between the UAV and object, and (3) image quality dependency due to UAV turbulence or various weather conditions.

We argue that developing object detection models which can adequately learn features invariant to these dependencies is key to substantially enhancing object detection accuracy in UAV-based perception. This requires curating UAV-based datasets that include images and metadata that carefully depict the full spectrum of correlations between the target scene on the ground and the camera on the UAV, whose imaging conditions constantly change as the UAV navigates the entire range of given operational requirements. Therefore, it is imperative to have a UAV-based object detection benchmark carefully curated with object poses, UAV positions, and weather information in the form of metadata for accurate model validation and verification.

Existing UAV-based benchmarks, such as VisDrone [9], UAVDT [10], Okutama-Action [11], and Standford Drone Dataset [12], provide limited metadata. Despite containing a wide variety of scenes captured using UAVs under different circumstances with various types of objects of interest, they do not provide a complete set of metadata, such as object poses and UAV positions, for each image in the datasets. This significant lack of information about how objects on the ground are projected through the camera lens as a function of UAV positions can lead to considerable limitations in learning about the objects in UAV-based images, as the appearances/attributes of the objects are subject to large variations.

To overcome the limitation resulting from lack of metadata, we introduce a large-scale UAV-based dataset, called Archangel, collected with comprehensive position and pose information by the DEVCOM Army Research Laboratory (ARL) (Fig. 1). Archangel comprises three sub-datasets: Archangel-Real [13], Archangel-Mannequin [13] and Archangel-Synthetic [14]. Archangel-Real comprises video sequences captured by a UAV flying at various altitudes and radii of rotation circles. It sets a group of real humans as targets, and each human is in one of three possible poses (i.e., stand, kneel or squat,¹ and prone) (Fig. 2). Similarly, Archangel-Mannequin sets a group of mannequins and different types of vehicles as targets. The imaging conditions for these two sub-datasets, such as UAV altitudes, ranges to targets, and object poses, are the same. Unlike Archangel-Real and Archangel-Mannequin, which were collected in real-world environments, Archangel-Synthetic was generated using the Unity game engine [15]. It includes a number of different virtual characters, who are in the same poses as described above and rendered with diverse illumination conditions. Archangel-Synthetic is designed for augmenting the other two real sub-datasets and for studying various issues tied to optimizing machine learning (ML) models using synthetic data, such as synthetic data augmentation and domain adaptation [16], with respect to UAV-based scene understanding.

In addition to collecting a new dataset, we further characterize Archangel using state-of-the-art (SoTA) object detection models, specifically the YOLOv5 family [17] which has five different levels of architectural complexity and is pre-trained on MS-COCO [8], a large-scale ground-based object detection dataset. In this paper, we focus on the YOLOv5 models with lower complexity (i.e., YOLOv5s, YOLOv5m, YOLOv5l) since they are able to run on computing resource typically available on small UAV platforms (Tab. 2). For each model, we evaluate its human detection performance using Archangel-Real. Since we programmed our UAVs to circle around the human targets at various altitudes and radii during data collection (Fig. 3), the detection accuracy can be compared across the whole range of UAV positions.

Furthermore, we optimize the pre-trained YOLOv5 models with different fine-tuning strategies using various hybrid combinations of subsets from Archangel-Mannequin and Archangel-Synthetic. These fine-tuning strategies have been designed to provide valuable guidelines for leveraging synthetic data in training ML models to boost their performance. A comprehensive performance comparison between the baseline and the optimized YOLOv5 models is presented to demonstrate how incorporating a combination of real and synthetic data can enhance detection performance across varying UAV positions (Secs. V and VI). One of the critical findings from the comparative performance analysis indicates that if the real and synthetic data used for fine-tuning is balanced with respect to the amount of data, a significant performance boost can be achieved even with a low-complexity model, such as YOLOv5s. Furthermore, the optimization based on the real and synthetic data is much more effective on infrequent object poses that are rarely seen in the original dataset for pre-training. For example, prone from Archangel is not often seen in MS-COCO [8], so the performance improvement on this pose is more evident.

¹The terms kneel and squat are used interchangeably throughout this paper.
This paper is extended from our previous preliminary studies [13], [14], where we mainly focused on introducing new datasets separately without extensive data analysis and characterization. As a result, the major scope of this paper is to extensively study the three sub-datasets jointly as a unified UAV-based benchmark with metadata for human detection. In summary, the contributions of this paper are as follows:

1) We present a unified Archangel dataset after substantially restructuring the three sub-datasets since our conference publications [13], [14], including additional labeling Archangel-Real and an expansion of the range of Archangel-Synthetic. Note that Archangel was not previously available for access due to incomplete labeling and restructuring. To the best of our knowledge, Archangel is the first UAV-based object detection dataset which contains real and synthetic sub-datasets captured with similar imaging conditions and includes an extensive set of metadata (e.g., object poses and UAV positions).

2) We conduct extensive data analysis on Archangel by jointly analyzing its three sub-datasets (Archangel-Synthetic, Archangel-Mannequin and Archangel-Real). In particular, we provide several important guidelines on exploiting real and synthetic data together to improve UAV-based object detectors.

II. RELATED WORK

A. UAV-BASED OBJECT DETECTION DATASETS

There is an increasing number of large-scale benchmarks for object detection, utilizing images captured with some fixed or moving cameras on the ground [7], [8], [18], [19], [20], [21], [22], yet relatively few datasets have been collected with UAVs. Moreover, these UAV-based object detection datasets all have their own limitations. For instance, VisDrone [9] is one of the major datasets for UAV-based object detection. It consists of images captured with UAVs in dozens of different scenarios, in which ten categories of objects were selected and carefully labelled. While it does have an advantage in data diversity, VisDrone does not provide any
metadata, such as UAV positions. In contrast to VisDrone, UAVDT [10], another well-known benchmark for UAV-based detection and tracking, provides three different kinds of UAV-specific metadata (i.e., weather condition, flying altitude and viewing angle) and a few object attributes such as vehicle category. However, the annotations for the metadata are coarse (i.e., 3 categories for each). Also, the dataset does not contain the human category. Unlike UAVDT, Okutama-Action [11] is composed of images from aerial views and contains humans in different human poses. However, it provides limited metadata regarding UAV altitudes and camera viewing angles. A212-Haze [23], which is the first real haze and object detection dataset with in-situ smoke measurements aligned to aerial imagery and is used in the recent UG²+ Challenge [24], provides UAV position metadata. However, A212-Haze does not include a synthetic subset, limiting its usage for facilitating studies on how to improve UAV-based object detectors by using synthetic data. More recently, DGTA [25] generates synthetic datasets associated with existing UAV-based object detection datasets and provides UAV position metadata for the generated datasets. Nevertheless, the existing UAV-based datasets they use, such as VisDrone [9], are still lack of metadata, limiting DGTA’s usage for precise model diagnosis.

There are two other types of datasets which are closely related to UAV-based object detection. First, datasets such as DOTA [26] include aerial images collected from satellites or aircraft. Although they are usually curated for remote sensing applications, detecting objects in such datasets also suffers from severe variations in the scale and orientation of the objects. Second, some UAV-based datasets are designed for certain vision tasks strongly associated with object detection. For instance, CARPK [27] is a large-scale car parking lot dataset designed for object counting. MOR-UV [28] is a large-scale moving object recognition dataset comprising videos captured by a UAV in various environments, such as urban areas and highways. Stanford Drone Dataset [12] is used for analyzing various object trajectories in the real world from the top-view. UAV123 [29] is used for low altitude UAV-based object tracking. Similar to Archangel, UAV123 also includes synthetic data generated by a photo-realistic simulator.

A comprehensive investigation of recent UAV-based datasets is shown in Tab. 1. Note that Archangel is the first ever UAV-based dataset collection not only containing both real and synthetic data but also providing an extensive set of metadata, including object poses and UAV positions.

B. UAV-BASED OBJECT DETECTION METHODS

With the rapid development of generic object detection methods [34] and the aforementioned UAV-based object detection benchmarks (Tab. 1), the detection accuracy of UAV-based object detectors has improved significantly over the past few years. In addition to common issues for generic object detection, UAV-based object detection has its own unique challenges [35]. In general, all of the challenges can be roughly divided into three categories. First, objects in UAV-based images are usually much smaller [36]. Therefore, many solutions have been proposed to address this problem to date. For example, Liu et al. [37] proposed HRDNet that fused information from both high- and low-resolution inputs to simultaneously preserve features of small objects and maintain computational costs. Similarly, Liu et al. [38] introduced a multi-branch and parallel structure (MPFPN) to extract more powerful features for tiny object detection. Besides the scale of an object, target objects in UAV-based datasets are usually crowded and sparsely distributed, reducing both the accuracy and efficiency of an object detector. Thus, Yang et al. [39] proposed ClusDet that performed object cluster proposal first before detecting objects. Finally, UAV-based datasets contain many UAV-specific nuisances [40], such as varying UAV altitudes and viewing angles. These nuisances cause tremendous variations in object appearances, causing degraded detection performance. To address this issue, Wu et al. [40] proposed to adopt adversarial training to learn domain-robust features from UAV-specific nuisances coarsely annotated by the authors. In this paper, we posit that UAV-based object detection can be further enhanced by providing UAV-based benchmarks with a set of fine-grained metadata, such as that contained in Archangel.

In addition to improving the detection accuracy, reducing the computational cost to achieve real-time on-board processing is also very important for UAV-based object detection approaches. One way to improve latency is to skip unnecessary computation. For instance, Ammour et al. [41] proposed to extract candidate regions of target objects first via over-segmentation. After that, only windows around the candidate regions were sent to the pre-trained CNN and linear SVM for feature extraction and classification. Another way of reducing computational overhead is to use more efficient one-stage object detectors, such as YOLO [17], [42], RetinaNet [43], CenterNet [44], and EfficientDet [45]. These one-stage object detectors directly classify and locate objects without generating region proposals, resulting in improved latency. As an example, Liu et al. [46] adapted the original network architecture of YOLO by making it more suitable for UAV-based object detection. In this paper, we also utilize YOLOv5 [17] for all the experiments due to the advantage of its low complexity (Tab. 2).

III. THE ARCHANGEL DATASET

The data collection process for Archangel is illustrated in Fig. 3 and a brief comparison of the three sub-datasets is provided in Tab. 3. In the following, we will go through the data collection process of each sub-dataset in detail.

A. ARCHANGEL-MANNEQUIN

1) TARGET OBJECTS

During this data collection, a group of mannequins were used as human surrogates primarily due to the safety guidelines...
TABLE 1. Comparison of recent UAV-based datasets. (1k = 1000).

| Name                | Yaks   | Year | #Clips | #Images | Resolution | SynReal | Human | Obc | Poses | UAV Fox | Lifting Cond. |
|---------------------|--------|------|--------|---------|------------|---------|-------|-----|-------|---------|----------------|
| Stanford Drone Dataset [12] | TF     | 2018 | 60     | 929.5k  | 1100x1904  | R       | ✓     | -   | -     | -       | -              |
| UAV-CI [28]         | OT     | 2018 | 125    | 112.6k  | 1130x720   | R       | ✓     | -   | -     | -       | -              |
| Obanja-Action [11]  | OD, AR | 2017 | 45     | 77.4k   | 3840x2160  | R       | ✓     | -   | -     | -       | -              |
| CARPF [27]          | OC     | 2017 |        | 1.4k    | 1130x720   | R       | ✓     | -   | -     | -       | -              |
| UAV/DT [10]         | OD, OT | 2018 | 100    | 80k     | 1024x540   | R       | ✓     | -   | -     | -       | -              |
| VALLDress [9]       | OD, OT | 2018 | 263    | 17.5k, static: 10k | various | R       | ✓     | -   | -     | -       | -              |
| DressSURF [9]       | FRD    | 2018 | 200    | 411.7k  | 1280x720   | R       | ✓     | -   | -     | -       | -              |
| AU-AIR [31]         | OD     | 2020 | 8      | 32.8k   | 1920x1080  | R       | ✓     | -   | -     | -       | -              |
| MOR-UAV [14]        | MOR    | 2020 | 30     | 10.9k   | various   | R       | ✓     | -   | -     | -       | -              |
| DOTA [26]           | DOTA   | 2021 |        | 11.3k   | various   | R       | ✓     | -   | -     | -       | -              |
| UAV-Human [53]      | AR, PE, PR, ATR | 2022 | 56-74k | various | various   | R       | ✓     | -   | -     | -       | -              |
| SeaHeroes [33]      | OD, OT | 2022 | 90     | 3.6k, MOT: 17.4k | various | R       | ✓     | -   | -     | -       | -              |
| AU2Life-Hare [23]   | IER, OD| 2022 |        | 1k      | 1645x1500  | R       | ✓     | -   | -     | -       | -              |
| DOTA [25]           | DOTA   | 2022 |        |         | 3840x2160  | S       | ✓     | -   | -     | -       | -              |

TABLE 2. Complexity of the YOLOv5 models [17] used in this study.

| Model      | #Parameters (M) | FLOPs (G) |
|------------|-----------------|-----------|
| YOLOv5n    | 3.1             | 4.3       |
| YOLOv5s    | 12.3            | 16.2      |
| YOLOv5m    | 35.3            | 49.1      |

FIGURE 3. Illustration of the data collection process for Archangel. For each data collection, a number of objects (real people, mannequins or virtual characters) on the ground were captured by a camera mounted on a UAV (real or simulated). Each of the objects was in one of the three defined poses (stand, kneel, and prone). The UAV circled around the objects at a predefined altitude and radii of rotation circles.

2) DATA COLLECTION

The imagery was captured using a contractor-built UAV equipped with an onboard electro-optical (EO) camera (ELP-USBFH01M-L21) with a 1920 x 1080 pixel array and a lens with approximately 120° field-of-view (FOV). The UAV camera was pitched forward by 45° relative to level flight. During the course of multiple UAV flights, the UAV operated over a wide range of altitudes and radii of rotation circles while keeping the camera pointed inward toward the targets and circling a central point. Both the altitude and radius of the rotation circle were varied from 15-50 meters in 5-meter increments. Since the target objects were stationary and the camera pitch angle was constant, the target objects were spread across different regions of the camera’s FOV, resulting in different view angles.

B. ARCHANGEL-SYNTHETIC

1) MOTIVATION

While Archangel-Mannequin provides valuable aerial imagery with pose and position metadata well suited for UAV-based human detection, the imaging conditions of this data collection were limited. Furthermore, Archangel-Mannequin did not incorporate some important factors, such as various human appearances/attributes, extended ranges of UAV altitudes and radii of rotation circles, and different illumination conditions. To overcome these restrictions, a large-scale synthetic imagery (i.e., Archangel-Synthetic) dataset containing multiple virtual humans in the same poses as that used in Archangel-Mannequin was generated using the Unity game engine [15] to augment Archangel.

2) DATA GENERATION AND LABELING

In the Unity-based simulation, a 3D scene is constructed using a terrain asset (i.e., background) and one or more target assets (i.e., virtual characters in different outfits and poses). For the current version of Archangel-Synthetic, we use only a simple terrain model (i.e., desert), but we plan to integrate more complex terrain models in future work.

For each target asset, we first created a Unity project (i.e., background) and one or more target assets (i.e., virtual characters in different outfits and poses). For the current version of Archangel-Synthetic, we use only a simple terrain model (i.e., desert), but we plan to integrate more complex terrain models in future work.
TABLE 3. Comparison of the three sub-datasets comprising Archangel.

| Archangel | Image Size | Targets | Altitude (m) | Radius (m) | Field of View | Camera Pitch Angle |
|-----------|------------|---------|--------------|------------|---------------|-------------------|
| Synthetic | 512×512    | Virtual characters | [5-80] increment by 5 | [5-80] increment by 5 | 22.5° | various |
| Mannequin | 1920×1080  | Mannequins, vehicles | [15-50] increment by 5 | [15-50] increment by 5 | 120° | 45° |
| Real      | 1304×978   | Real people       | [5-80] increment by 5 | [20-50] increment by 5 | 45° | 22.5°, 45°, 67.5° |

To synthesize images and annotations from the virtual 3D environment constructed above, we used an open source software asset, Image Synthesis for Machine Learning [47]. Specifically, the software produced an image segmentation mask where each target object in a synthetic image was assigned a unique scalar value. To generate the bounding box annotations for each target, a Python script was used to parse each segmentation mask, identify each target object, and measure the center, width, and height of the tightest bounding box encompassing the target. Additionally, the target category, the camera position, the target orientation relative to the camera, the camera-to-target distance, the camera pitch angle, and the number of pixels inside the segmentation mask were recorded in a single JavaScript Object Notation (JSON) file for each trial.

3) PROPERTIES
Archangel-Synthetic includes mountainous desert terrain and eight different virtual characters, each in three different poses (i.e., stand, squat, and prone). In a single trial of synthetic data generation (i.e., a virtual character with a certain pose), both the altitude of the camera and the radius of the rotation circle were varied from 5-80 meters in 5-meter increments. Additionally, the camera viewing angle relative to the character was varied from 0°-358° in 2° increments and four different sun angles were simulated. This resulted in over 4.4M images included in Archangel-Synthetic. Each image contains 512×512 pixels with horizontal and vertical fields-of-view of 22.5°.

C. ARCHANGEL-REAL
1) SYSTEM DESIGN
The dataset was collected with an ARL-designed UAV platform called the Dawn Dove (D2). The D2 is a re-configurable UAV, with the ability to shift the center of gravity by adjusting arm angles, arm placement, battery, sensor payload, and on-board processor location. It is composed of a combination of 3D printed polyethylene terephthalate glycol (PETG) and carbon fiber infused nylon, and traditional carbon fiber parts. It can carry various types of sensor payloads and on-board processors. It has an approximate payload capacity of 1.5 lbs and an approximate flight time of 8 minutes. For this data collection, the sensor payload consisted of a UI-3250ML-C-HQ EO camera with an Edmund Optics 6mm/F1.4 lens and a FLIR Boson 640 8.7 mm IR camera. The cameras were co-located on the front of the D2 and the EO camera’s image was cropped to match the FOV of the FLIR Boson (50° HFOV). The on-board processor was an NVIDIA Xavier NX with a 1 TB NVMe SSD for additional data storage.

2) DATA COLLECTION
In this dataset, the targets consisted of real people wearing civilian clothing in three different poses: stand, kneel, and prone. The data collection process involved having the D2 fly circles at radii ranging from 20-50 meters, at intervals of 5 meters, and at altitudes ranging from 15-50 meters, at intervals of 5 meters. The camera angle relative to level flight was manually adjusted between flights from −22.5°, −45°, and −67.5° to ensure the targets remained within the FOV of the cameras. In total, 52 circles were flown around the targets. To fly the circles, custom Robot Operating System (ROS) based autonomy code was used, along with a custom Python-based graphical user interface (GUI), which communicated with the UAV. From the ground control station (GCS), the target GPS location, circle radius, altitude, maximum velocity, and file name were entered into the GUI. Once the circle parameters were entered, the D2 was manually armed, launched, and switched over to “offboard mode” which passed control of the UAV to the GCS. The GCS then commanded the UAV to perform the autonomous circle. Once complete, the next circle’s parameters were entered into the GUI and sent to the UAV while still in the air. This was repeated each flight until the UAV had to be brought back down to replace the battery. In addition to stationary targets, a few circles were also flown where the people walked, jogged, crawled, and waved. Note that Archangel-Real involves human subjects as UAV-based detection instances. However, an Institutional Review Board (IRB) approval was exempted since one cannot identify individuals in the dataset.

D. IMPORTANCE OF THE CAMERA PARAMETERS
Before moving on to the data analysis section, we want to highlight the importance of revealing the camera parameters used in the data collection. Note that the scale of human instances in UAV-based object detection datasets, such as Archangel, is strongly influenced by the camera parameters used in the data collection, including FOV, pixel-array size,
and pitch angles, in addition to the UAV altitude and radius of rotation circle. Hence, the detection results can vary greatly when using different camera parameters. However, all the conclusions derived from the following data analysis of Archangel can still be applied to other UAV-based datasets using different camera parameters through extrapolation.

That is, the performance gap can be easily calibrated by adjusting the scale of human instances if the camera parameters and the original object size are known a priori.

IV. EXPERIMENTAL SETUP
A. OVERVIEW
In this paper, we designed a series of experiments based on the flow shown in Fig. 4. In brief, for each experiment, we selected one of the three pre-trained YOLOv5 models (Tab. 2) and fine-tuned the model on varying amounts of UAV-based real (i.e., Archangel-Real) and synthetic (i.e., Archangel-Synthetic) data. We then evaluated the model on a sequestered UAV-based dataset (i.e., Archangel-Real). Based on the results, the designed experimental flow can provide valuable insights into optimizing UAV-based object detectors with hybrid sets of real and synthetic data.

B. DATASETS
Note that acquiring the best performance for an UAV-based object detector is not the primary purpose of this study. Thus, we subsampled each of the three sub-datasets of Archangel to explore optimal strategies for fine-tuning or evaluating models:

1) Archangel-Mannequin: The dataset consists of video clips collected in 11 UAV flight trials. In this paper, we carefully split the dataset into two subsets so that each covered the whole range of the UAV positions covered during the entire data collection. The video clips collected in 6 of the 11 trials (i.e., Trial-5, 6, 8, 9, 10, 11) were used for evaluating models. The rest (i.e., Trial-1, 2, 3, 4, 7) were used for fine-tuning models. All the video clips were uniformly subsampled at 3 fps. This resulted in two sets of frames, Arch-Mann-FT37, containing 6.7k frames for fine-tuning models, and Arch-Mann-Eval, containing 11.2k frames for evaluating models. Arch-Mann-FT37 is named based on the amount of data it has compared to the entire Archangel-Mannequin in terms of percentage (i.e., 37%).

2) Archangel-Real: Similarly, we uniformly subsampled the video clips in Archangel-Real at 1 fps. This resulted in a set of frames, named as Arch-Real-Eval, containing 4.1k frames for evaluating models.

3) Archangel-Synthetic: Only one virtual character in all of the three poses was used. For each UAV position, only one of the four sun angles was randomly selected. Additionally, instead of using all the UAV positions, we uniformly sampled images across each rotation circle in 60° increments. This resulted in a set of images, named as Arch-Syn-FT, containing 4.6k images for fine-tuning models.

Each of the three sub-datasets has its own unique usage in this study. In general, Archangel-Real serves as the primary UAV-based benchmark for measuring detection accuracy. Archangel-Mannequin can be viewed as the real UAV-based fine-tuning dataset for adapting the detection models to the target UAV-based domain. Although it includes mannequins instead of real humans, fine-tuning on this dataset is shown to be effective in the following sections. Archangel-Synthetic, on the other hand, is used as the synthetic version of the UAV-based fine-tuning dataset, which can be combined with Archangel-Mannequin to further optimize the models.

C. EVALUATION
We utilized standard AP50, the average precision with an IOU (Intersection of Union) threshold of 0.5, as the metric to measure the performance of each object detector. Moreover, AP50 was computed for each pose respectively. As more than one pose may exist in a single image, to obtain the performance for only one certain pose, the other two poses were ignored during the evaluation process.

D. IMPLEMENTATION DETAILS
The official repository of YOLOv5 [17] was used for both fine-tuning and evaluating models. If not specified otherwise, the default hyperparameters were adopted. The input images were rescaled (i.e., imsz=1280) first before being fed into all the models. During fine-tuning, the backbone for each model was frozen (i.e., freeze=10) to prevent the model from easily over-fitting. We fine-tuned each model for 20 epochs with a batch size of 16 on a server with 4 NVIDIA GeForce RTX 2080 Ti GPUs. During the evaluation, we set the confidence threshold to be 0.05.

V. RESULTS
A. THE PERFORMANCE OF PRE-TRAINED MODELS
To begin with, we evaluated the three pre-trained YOLOv5 models on Arch-Mann-Eval and Arch-Real-Eval. The models were pre-trained on MS-COCO, a representative ground-based dataset. The results are shown in Fig. 5. From the results, we can gain several useful insights on UAV-based object detection. First, larger pre-trained models achieved...
better accuracy across all the evaluation datasets and poses. One possible reason for this is that larger models, compared with smaller ones, can explore better and find more powerful features for classification and detection [48]. Although it implies that we can get higher accuracy by using larger pre-trained models, such larger models may not fit well on small UAV platforms with computational constraints.

Another trend we can observe is that the pre-trained models had much better accuracy on stand. It is mainly because the dataset used for pre-trained models (i.e., MS-COCO) contains significantly more human instances in stand (i.e., 84.53%) [49]. In other words, it is impractical to directly use detectors pre-trained on standard datasets, especially in unusual scenarios where we need to detect people in uncommon positions, such as search and rescue in disaster relief. It is worth mentioning that the pre-trained YOLOv5 models performed much worse on Arch-Mann-Eval than on Arch-Real-Eval. That is mainly because Arch-Mann-Eval contains some other objects, such as traffic cones and fiducials, which are easily misclassified as humans when captured by the pre-trained detectors at high altitudes [13].

### B. FINE-TUNING MODELS ON A REAL UAV-BASED DATASET: ARCH-MANN-FT37

As we have discussed, most ground-based datasets used to fine-tune models usually lack UAV-specific samples for the models to learn from, such as human instances in non- standing positions captured from various camera viewing angles and altitudes. Therefore, we allowed the pre-trained YOLOv5 models to acquire such knowledge by fine-tuning the models on Arch-Mann-FT37. Moreover, given the significant challenges of collecting and annotating UAV-based datasets [9] and the lack of existing large-scale UAV-based object detection benchmarks, we explored the idea of fine-tuning models in the small-data regime [50]. More precisely, we subsampled the original fine-tuning dataset and created several smaller subsets for fine-tuning, which contained much less data (i.e., Arch-Mann-FT20, Arch-Mann-FT10, Arch-Mann-FT5 and Arch-Mann-FT2). We followed the same naming strategy as Arch-Mann-FT37 for the extra fine-tuning datasets.

The results are presented in Fig. 6. We would like to highlight the importance of having an evaluation dataset with different characteristics from the fine-tuning dataset. As we fine-tuned the models on data from Arch-Mann-FT37, we could improve their detection accuracy on a similar evaluation dataset such as Arch-Mann-Eval. However, the models fine-tuned on too much data from Arch-Mann-FT37 tended to perform worse on Arch-Real-Eval. We argue that this was because the models started to learn certain dataset-specific features from the fine-tuning dataset, adversely affecting the models’ generalization capability to unseen datasets, such as the evaluation dataset in our case. In practice, ML models embedded into UAVs are usually deployed to new environments unseen during training. Thus, in the following experiments, we chose to fine-tune models on data from Arch-Mann-FT37 and Arch-Syn-FT but evaluate them on Arch-Real-Eval.

### C. FINE-TUNING MODELS ON BOTH REAL AND SYNTHETIC DATASETS: ARCH-MANN-FT37 AND ARCH-SYN-FT

One of the significant advantages of Archangel is that it contains both real and synthetic subsets acquired from similar imaging conditions in data collections and synthetic rendering, respectively. Hence, investigating the effect of augmenting the original UAV-based fine-tuning datasets with UAV-based synthetic data is another major topic for this study. To do so, we fine-tuned the pre-trained YOLOv5 models on Arch-Syn-FT. Additionally, a well-known concern about learning from synthetic data is that, compared with real data, synthetic data usually contains much less variations in appearances/attributes of objects or structures of scenes [51]. Therefore, instead of fine-tuning models only on Arch-Syn-FT, we also explored the idea of multi-source learning [52], constructing multiple hybrid fine-tuning datasets by directly merging Arch-Syn-FT with all the subsets of Arch-Mann-FT37 used in the previous experiment.

The results are shown in Fig. 7. For prone, a rarely seen pose in the pre-training dataset, the detection accuracy of the pre-trained models was very low. After fine-tuning the models on the various hybrid subsets of Arch-Syn-FT and Arch-Mann-FT37, the detection accuracy continually increased with few exceptions. For stand, the pre-trained models performed much better than they did for prone as expected, but fine-tuning on the hybrid sets of the real and synthetic data still provided much improvement over the pre-trained models, as clearly observed from the detection accuracy of YOLOv5n6. Similar observations can be made for kneel.

We now compare the results shown in Fig. 6 with the ones shown in Fig. 7 to highlight the effect of adding the synthetic data to the fine-tuning dataset based only on the subsets of Arch-Mann-FT37. The results are shown in Fig. 8. For prone, we can observe significant performance improvement across all the models and hybrid subsets for fine-tuning after introducing the synthetic data into the fine-tuning datasets, compared with the results of fine-tuning on data from Arch-Mann-FT37 only. One explanation for the above finding is that most of the human instances in the pre-training dataset are in certain upright positions, including stand and kneel. Therefore, adding synthetic characters in prone captured from
Y.-T. Shen et al.: Archangel: A Hybrid UAV-Based Human Detection Benchmark

Figure 6. AP50 of the YOLOv5 models fine-tuned on the different subsets of Arch-Mann-FT37. Each model was evaluated on Arch-Mann-Eval (top) and Arch-Real-Eval (bottom).

Figure 7. AP50 of the YOLOv5 models fine-tuned on the various hybrid sets constructed by Arch-Syn-FT and the different subsets of Arch-Mann-FT37.

Various camera viewing angles and UAV altitudes to the fine-tuning dataset can effectively aid the models to learn how to detect humans in prone. More in-depth analysis of this issue will be discussed in the ablation study (Sec. VI).

Additionally, Arch-Syn-FT, which includes only one virtual character and one type of background, might be too simple for larger models, such as YOLOv5s6 and YOLOv5m6 in our case, to learn from, causing them to overfit to the fine-tuning dataset. A proof of this assumption was that using Arch-Syn-FT along with the subsets of Arch-Mann-FT37 to fine-tune YOLOv5s6 and YOLOv5m6 significantly decreased their performance on stand and kneel, especially when we included fewer data from Archangel-Mann-FT37 (Fig. 8). On the other hand, fine-tuning YOLOv5n6 on Arch-Syn-FT along with the subsets of Arch-Mann-FT37 did not have such a negative effect. As a result, in the following ablation study, we focused on analyzing the results of YOLOv5n6 once we included Arch-Syn-FT in the fine-tuning dataset.

VI. ABLATION STUDY
A. ADJUSTING THE SIZE OF THE SYNTHETIC DATASET: ARCH-SYN-FT

We have demonstrated the effects of directly combining the various subsets of Arch-Mann-FT37 with the same set of synthetic data (i.e., Arch-Syn-FT) for fine-tuning models (Fig. 7 and 8). In this section, we are interested in further exploring the outcome of using different amounts of synthetic data to fine-tune models. Notably, we aim to investigate whether a “balanced” fine-tuning dataset, which contains the same amount of real and synthetic data, is better than its “unbalanced” counterpart, in which the ratio of the synthetic data to the real data varies due to the use of a fixed set of the synthetic data across all the hybrid fine-tuning datasets.

To achieve this goal, for each real fine-tuning dataset (i.e., Arch-Mann-FT37, Arch-Mann-FT20, Arch-Mann-FT10, Arch-Mann-FT5, and Arch-Mann-FT2), the corresponding amount of data was randomly sampled from Arch-Syn-FT to match the real fine-tuning dataset. Namely, if the real fine-tuning dataset was smaller than Arch-Syn-FT, a subset of Arch-Syn-FT was randomly selected and combined with the real fine-tuning dataset to form a balanced fine-tuning dataset. Similarly, if the real fine-tuning dataset was larger, a random subset of Arch-Syn-FT was duplicated before the combination. We denote the synthetic fine-tuning dataset as Arch-Syn-FT-B if the aforementioned data balancing procedure has been conducted.

The results are shown in Figs. 9, 10 and 11. Comparing Fig. 8 with Fig. 10, we can observe that the negative effect of fine-tuning larger models (i.e., YOLOv5s6 and YOLOv5m6) on hybrid fine-tuning sets largely decreases or
even diminishes. Moreover, the improvement becomes more significant, especially for \texttt{YOLOv5n6}. From Fig. 11, it is shown that fine-tuning \texttt{YOLOv5n6} on the balanced hybrid sets of the real and synthetic data can always be at least on par with the best performance setting for fine-tuning. These findings indicate that the proportion of each data to the whole fine-tuning dataset significantly impacts the model’s fine-tuning performance.

**B. LEAVING ONE POSE OUT FROM THE SYNTHETIC DATASET: ARCH-SYN-FT**

We have claimed that fine-tuning models with synthetic human instances are particularly important for \texttt{prone}, a pose rarely seen in the original training data. In this section, we would like to provide more evidence to support this statement. Specifically, we used \texttt{YOLOv5n6} for this set of experiments since it showed less sign of overfitting when fine-tuning on \texttt{Arch-Syn-FT} (Fig. 7). Additionally, the technique of dataset balancing was adopted due to its positive effect on fine-tuning models (Fig. 9 and 10). We followed a similar procedure to generate each hybrid fine-tuning dataset, except that each time one of the poses, \texttt{stand}, \texttt{kneel}, or \texttt{prone}, was excluded in advance from \texttt{Arch-Syn-FT}, resulting in a “leave-one-pose-out” fine-tuning dataset, \texttt{Arch-Syn-FT-NoSt}, \texttt{Arch-Syn-FT-NoKn}, or \texttt{Arch-Syn-FT-NoPr}, respectively.

The results are presented in Fig. 12. In general, the detection accuracy of \texttt{stand} and \texttt{kneel} did not obviously change when we removed any one of the poses from the synthetic fine-tuning dataset. However, the performance of \texttt{prone} degraded drastically when we excluded the virtual characters in \texttt{prone} from the fine-tuning dataset, which strongly supports our earlier claim.

**C. PRECISE MODEL DIAGNOSIS: PERFORMANCE COMPARISON ON THE ALTITUDE/RADIUS GRID**

So far, we have demonstrated that the detection accuracy of the pre-trained \texttt{YOLOv5} models on the UAV-based evaluation dataset can be considerably boosted by fine-tuning the models on the hybrid sets of UAV-based real and synthetic data. For example, we have shown that the AP50 of the pre-trained \texttt{YOLOv5n6} can be increased by about 30 in AP...
FIGURE 11. AP50 comparison of YOLOv5n6 fine-tuned on the different hybrid sets of Arch-Syn-FT, Arch-Syn-FT-B, and the different subsets of Arch-Mann-FT.

FIGURE 12. AP50 of YOLOv5n6 fine-tuned on the various balanced hybrid sets constructed by the three leave-one-pose-out synthetic fine-tuning datasets (i.e., Arch-Syn-FT-NoSt-B, Arch-Syn-FT-NoKn-B and Arch-Syn-FT-NoPr-B) and the different subsets of Arch-Mann-FT.

FIGURE 13. AP50 comparison on the altitude/radius grid of YOLOv5n6 fine-tuned on the different hybrid sets of Arch-Mann-FT2, Arch-Syn-FT and Arch-Syn-FT-B.
value by fine-tuning it on a joint set of Arch-Mann-FT2 and Arch-Syn-FT-B (Fig. 11). In this section, we further analyze this particular example with the complete information about the UAV positions over the altitude/radius grid as shown in Fig. 13. Our goal is to give an idea of how to utilize the metadata provided by Archangel to diagnose problems with a UAV-based object detection model.

The results are presented in Fig. 13. In this figure, we can clearly observe how the pre-trained YOLOv5n6’s performance gradually progresses with the different fine-tuning datasets, from the real-data-only dataset to the unbalanced hybrid dataset to the balanced hybrid dataset. Initially, the pre-trained YOLOv5n6 performs fairly well at the low altitudes but fails at the high altitudes due to the curse of the pre-training dataset, which is composed mostly of ground-based human instances. After being fine-tuned on Arch-Mann-FT2, the fine-tuned YOLOv5n6 gets much better at detecting human instances from a relatively higher altitude or larger circle radii. However, the detection accuracy of the human instances at close range decreases considerably. We argue that this is mainly because the image resolution of Arch-Mann-FT2 (i.e., 1920 × 1080) is much larger than that of the evaluation dataset (i.e., 1304 × 978) and the bounding boxes contained in Arch-Mann-FT2 are generally small. In other words, fine-tuning YOLOv5n6 on Arch-Mann-FT2 inevitably causes a bias on the model toward detecting tiny objects. In contrast, this negative effect does not occur when we fine-tune the model on the hybrid set of Arch-Mann-FT2 and Arch-Syn-FT (or Arch-Syn-FT-B). We believe that this is because our synthetic dataset covers an extended range of UAV positions over the grid so as to cause less bias on the fine-tuning process.

Based on the results shown in Fig. 13, to further improve the model’s detection accuracy, we can focus on improving the model’s performance at high altitudes and large circle radii, or the performance of non-standing positions, such as kneel and prone. Notably, we find that the performance improvement of kneel is surprisingly small to high altitudes and large circle radii, which requires a deeper investigation in future work.

**VII. DISCUSSION AND FUTURE DIRECTION**

With the series of experiments presented in Sec. V and VI, we have demonstrated the distinctive value of the Archangel dataset. Particularly, we have clearly illustrated how to utilize the dataset’s metadata to evaluate and diagnose the UAV-based object detectors on the altitude/radius grid. Moreover, we have systematically analyzed how to involve both real and synthetic data within the UAV-based fine-tuning process. Such fundamental studies of UAV-based perception had not been achieved until we curated Archangel.

Despite having all these merits, Archangel is still in its early stage and has much room for improvement and development. In the following, we suggest a few possible future directions regarding Archangel:

**A. DIRECT EXTENSION OF THIS STUDY**

Many results shown in Sec. V and VI imply that increasing the data diversity of Archangel is one of the most promising future research directions. For instance, since we have demonstrated that fine-tuning models with virtual characters in unusual poses is particularly effective, we can include more atypical poses into Archangel to further explore this phenomenon, which is crucial especially in search and rescue scenarios where finding people in severe physiological states is the priority. Additionally, we can diversify the appearances/attributes of either the real or synthetic human instances in the dataset, investigating whether this will resolve the issue of overfitting as we have discussed earlier. For the same purpose, we can increase the diversity of components beyond the foreground objects, such as the real and synthetic backgrounds included in Archangel. Finally, we can extend Archangel to include more object categories, such as various types of vehicles, which frequently exist in other UAV-based object detection datasets (Tab. 1) so that Archangel can be used in conjunction with those datasets.

Next, exploring more sophisticated fine-tuning strategies is another potential direction to extend this work. In this study, we have demonstrated that the performance of the pre-trained SoTA object detector can be boosted considerably by fine-tuning the model on a balanced UAV-based fine-tuning dataset constructed by directly merging a real subset and a synthetic subset. Nevertheless, it is worth exploring if there is a better strategy for sampling each subset or merging the two subsets. For instance, to build up a balanced fine-tuning dataset, instead of randomly selected samples from the synthetic fine-tuning dataset, we may do the sampling based on certain distance measurements.

**B. UAV-BASED VISUAL REPRESENTATION LEARNING WITH METADATA**

In this study, we used the position and pose metadata provided by Archangel only for accurate model evaluation and diagnosis. However, we believe that there is a huge potential to utilize such metadata during training for better visual representation learning. A notable example for this is NDFT [40], where the authors exploited adversarial training with the coarse metadata labeled by themselves to enhance the robustness of the learned features for UAV-based object detection. We expect that such a framework will benefit greatly from the extensive metadata provided by Archangel. Beyond UAV-based perception, in medical [53] and underwater [54] imaging, it has also been shown that dataset metadata is useful for learning visual representation with self-supervised learning or contrastive learning.

**C. UAV-BASED SYNTHETIC DATA GENERATION AND AUGMENTATION**

We have found that fine-tuning larger models with the synthetic data that we generated often causes the issue of overfitting. This issue might be mitigated by directly
synthesizing more diverse images. However, as we have discussed, there is usually a huge domain gap between the synthetic data and real data in terms of object appearances/attributes and scene structures, which may not be solved by simply increasing the number of synthetic images. Hence, addressing this domain gap issue within the scope of UAV-based perception is another important future research direction for Archangel. Possible solutions include: (1) jointly training an object detector with a generative model which transforms synthetic images to be more visually realistic [55, 56], and (2) formulating the process of synthetic data generation as a learning problem to synthesize scene structures better matching real-world scene distributions [57].

VIII. CONCLUSION

In this paper, we introduce a unique UAV-based object detection dataset, Archangel, to encourage the community to continue developing more effective UAV-based object detection approaches with dataset metadata and synthetic data. A comprehensive study is carefully designed to show how to utilize Archangel to fully optimize a state-of-the-art object detector with a hybrid fine-tuning dataset comprising both real and synthetic data. Additionally, we also demonstrate the huge benefit of leveraging the dataset metadata during model evaluation by comparing the performance of the model across the different object poses and UAV positions. As we have discussed, although there is still much room for improvement, we hope that Archangel is useful for the broader machine learning community and can lead to future advances in the area of UAV-based perception.

ACKNOWLEDGMENT

The authors would like to thank the U.S. Army Artificial Innovation Institute (A2I2), for supporting the annotation of Archangel.

REFERENCES

[1] M. Erdelj and E. Natalello, “UAV-assisted disaster management: Applications and open issues,” in Proc. Int. Conf. Comput. Netw. Commun. (ICNC), Feb. 2016, pp. 1–5.
[2] A. Chirik, H. Touati, H. Snoussi, and F. Kamoun, “UAV-based surveillance system: An anomaly detection approach,” in Proc. IEEE Symp. Comput. Commun. (ISCC), Jul. 2020, pp. 1–6.
[3] M. Yaqot and B. C. Menezes, “Unmanned aerial vehicle (UAV) in precision agriculture: Business information technology towards farming as a service,” in Proc. 1st Int. Conf. Emerg. Smart Technol. Appl. (eSmart'19), Aug. 2021, pp. 1–7.
[4] M. A. Khan, W. Ectors, T. Bellemans, Y. Ruichek, A.-U.-H. Yasar, D. Janssens, and G. Wets, “Unmanned aerial vehicle-based traffic detection, and visual relationship detection at scale,” Int. J. Comput. Vis., vol. 128, no. 7, pp. 1956–1981, Jul. 2020.
[5] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2012, pp. 3354–3361.
[6] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, “Object365: A large-scale, high-quality dataset for object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit. (ICCV), Oct. 2019, pp. 8429–8438.
[7] P. Narayanan, X. Hu, Z. Wu, M. D. Thielke, J. G. Rogers, A. V. Harrison, J. A. D’Agostino, J. D. Brown, L. P. Quang, J. R. Uplinger, H. Kwon, and Z. Wang, “A multi-purpose realistic haze benchmark with quantifiable haze levels and ground truth,” 2022, arXiv:2206.06427.
[8] M. Mueller, N. Smith, and B. Ghanem, “A benchmark and simulator for UAV tracking,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 445–461.
[30] I. Kalra, M. Singh, S. Nagpal, R. Singh, M. Vatsa, and P. B. Sujit, “DroneSURF: Benchmark dataset for drone-based face recognition,” in Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), May 2019, pp. 1–7.

[31] I. Boscian and E. Kayacan, “AU-ALL: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2020, pp. 8504–8510.

[32] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, “UAV-human: A large benchmark for human behavior understanding with unmanned aerial vehicles,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 16261–16270.

[33] L. A. Varga, B. Kiefert, M. Messener, and A. Zell, “SeaDrones/See: A maritime benchmark for detecting humans in open water,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2022, pp. 3686–3696.

[34] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, “A survey of modern deep learning based object detection models,” Digit. Signal Process., vol. 126, Jun. 2022, Art. no. 103514.

[35] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, “Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey,” IEEE Geosci. Remote Sens. Mag., vol. 10, no. 1, pp. 91–124, Mar. 2022.

[36] X. Yu, Z. Han, Y. Gong, N. Jan, J. Zhao, Q. Ye, J. Chen, Y. Feng, B. Zhang, X. Wang, Y. Xin, J. Liu, M. Mao, S. Xu, B. Zhang, S. Han, C. Gao, W. Tang, L. Jin, M. Hong, Y. Yang, S. Li, H. Luo, Q. Zhao, and H. Shi, “The 1st tiny object detection challenge: Methods and results,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2020, pp. 315–323.

[37] Z. Liu, G. Gao, L. Sun, and Z. Fang, “HRDNet: High-resolution detection network for small objects,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2021, pp. 1–6.

[38] Y. Liu, F. Yang, and P. Hu, “Small-object detection in UAV-captured images via multi-branch parallel feature pyramid networks,” IEEE Access, vol. 8, pp. 145740–145750, 2020.

[39] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, “Clustered object detection in aerial images,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8310–8319.

[40] Z. Wu, K. Aares, P. Narayanan, H. Xu, H. Kwon, and Z. Wang, “Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1201–1210.

[41] N. Aamour, H. Alhichri, Y. Bazi, B. Benjdira, N. Alajlan, and M. Zuair, “Deep learning approach for car detection in UAV imagery,” Remote Sens., vol. 9, no. 4, p. 312, Mar. 2017.

[42] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal speed and accuracy of object detection,” 2020, arXiv:2004.10934.

[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 318–327, Feb. 2020.

[44] K. Duan, S. Bui, L. Xie, H. Qi, Q. Huang, and Q. Tian, “CenterNet: Keypoint triplets for object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6567–6577.

[45] M. Tan, R. Pang, and Q. V. Le, “EfficientDet: Scalable and efficient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10778–10787.

[46] M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, and C. Piao, “UAV-YOLO: Small object detection on unmanned aerial vehicle perspective,” Sensors, vol. 20, no. 8, p. 2238, Apr. 2020.

[47] ML-Imagesynthesis, U3DC, Unity Technol., San Francisco, CA, USA, 2017.

[48] A. Brutzkus and A. Globerson, “Why do larger models generalize better? A theoretical perspective via the XOR problem,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 822–830.

[49] Y. Huang, B. Sun, H. Kan, J. Zhuang, and Z. Qin, “FollowMeUp sports: New benchmark for 2D human keypoint recognition,” in Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV). Cham, Switzerland: Springer, 2019, pp. 110–121.

[50] J. Bornschein, F. Visin, and S. Osindero, “Small data, big decisions: Model selection in the small-data regime,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 1035–1044.

[51] V. Seib, B. Lange, and S. Wirtz, “Mixing real and synthetic data to enhance neural network training - a review of current approaches,” 2020, arXiv:2007.08781.

[52] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, “Domain generalization: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 4, pp. 4396–4415, Apr. 2023.

[53] Y. N. T. Vu, R. Wang, N. Balachandar, C. Liu, A. Y. Ng, and P. Rajpurkar, “MedAug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation,” in Proc. Mach. Learn. Healthcare Conf., 2021, pp. 755–769.

[54] T. Yamada, M. Massot-Campos, A. Prügel-Bennett, S. B. Williams, O. Pizarro, and B. Thornton, “Leveraging metadata in representation learning with georeferenced seafloor imagery,” IEEE Robot. Autom. Lett., vol. 6, no. 4, pp. 7815–7822, Oct. 2021.

[55] L. Liu, M. Muehly, J. Deng, T. Pfister, and L.-J. Li, “Generative modeling for small-data object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6072–6080.

[56] Y.-T. Shen, H. Lee, H. Kwon, and B. S. Bhattacharyya, “Progressive transformation learning for leveraging virtual images in training,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2023, pp. 835–844.

[57] A. Kar, A. Prakash, M.-Y. Liu, E. Cameracci, J. Yuan, M. Rusiniak, D. Acuna, A. Torralba, and S. Fidler, “Meta-Sim: Learning to generate synthetic datasets,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 4550–4559.

Y.-T. Shen (Member, IEEE) received the B.S. degree in electrical engineering and the M.S. degree in electronics engineering from the National Taiwan University, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Maryland, College Park. His research interests include computer vision, machine learning, and hardware/software co-design.

Yaesop Lee received the bachelor’s degree in electrical engineering from Soong University, Seo, South Korea, in 1984, and the master’s degree in ENTS from the University of Maryland, College Park, where she is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. Her research interests include embedded computer vision and real-time image/signal processing.

Heesung Kwon received the B.Sc. degree in electronic engineering from Soang University, Seoul, South Korea, in 1984, and the M.S. and Ph.D. degrees in electrical engineering from the State University of New York at Buffalo, in 1995 and 1999, respectively. He is a Senior Researcher and a Team Lead with the DEVCOM Army Research Laboratory (ARL). He served as one of the government leads of an ARL Collaborative Research Program—the Internet of Battlefield Things (IoBT). He has published over 150 journal articles, book chapters, and conference papers on various topics. He was a co-recipient of the Best Paper Award at the Army Science Conference, in 2004; and the Best Paper Runner-Up Award at the IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS 2016). He has been on technical program committee for various conferences and workshops relevant to image/video analytics and machine learning. He served as an Associate Editor for IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS.
DAMON M. CONOVER received the Ph.D. degree from George Washington University, in 2015. He is an Electrical Engineer with the Intelligent Perception Branch, DEVCOM Army Research Laboratory (ARL). His research focuses on 3-D geospatial data processing and visualization, simulation and synthetic data generation, and robotic systems. He is particularly interested in the intersection of geospatial information and robotics for high-level planning.

SHUJRA S. BHATTACHARYA (Fellow, IEEE) received the Ph.D. degree from the University of California at Berkeley. He is a Professor with the Department of Electrical and Computer Engineering, University of Maryland, College Park. He holds a joint appointment with the University of Maryland Institute for Advanced Computer Studies (UMIACS). He is also affiliated with the Maryland Crime Research and Innovation Center (MCRIC). He also holds a part-time visiting position as the Chair of excellence in design methodologies and tools with Institut National des Sciences Appliquées (INSA), Rennes, France. He has held industrial positions as a Researcher with the Hitachi America Semiconductor Research Laboratory and a Compiler Developer with Kuck and Associates. From 2015 to 2018, he was a part-time Visiting Professor with the Department of Pervasive Computing, Tampere University of Technology (currently Tampere University), Finland, as part of the Finland Distinguished Professor Program. He has also held visiting research positions with the U.S. Air Force Research Laboratory (AFRL).

NIKOLAS VALE is a Mechanical Engineer with the DEVCOM Army Research Laboratory (ARL). Autonomous Sensing and Integration Branch. His focuses are on designing, constructing, programming, and operating unmanned aerial vehicle (UAV) research platforms. He is particularly interested in developing modular, reconfigurable hardware, and software systems, for the purposes of data collection, experimentation, and prototyping.

JOSHUA D. GRAY is a Mechanical Engineer with iibertek, Inc. working with the DEVCOM Army Research Laboratory (ARL), Autonomous Sensing and Integration Branch. His focuses are on designing, constructing, and operating unmanned aerial vehicle (UAV) research platforms. He is particularly interested in developing small and multi-rotor UAVs, for experimentation and data collection.

G. JEREMY LEONG received the Ph.D. degree in applied chemistry (catalyst science and engineering) from the Colorado School of Mines. He is a Program Manager with DTRA’s Counter-WMD Advanced Research Division. Prior to joining DTRA, he spent 13 years in energy research, most recently six years with the U.S. Department of Energy, where he oversaw several programs in fuels and chemicals, advanced materials, and modeling and simulation/high performance computing for multiple directorates. In the past eight years, he managed a total Federal Applied Research and Development Budget of over $180M with numerous recognitions for successful technology transitions and deployments. He has authored more than 30 peer reviewed manuscripts, internal assessments, and congressional reports. He has also served on numerous technical panels and presented more than two dozen conference talks on modeling and simulation, software development, and the chemistry of materials and their practical applications.

KENNETH EVENSEN has served in various levels of command in the United States Army, for 27 years. Notably, he was the Program Manager of developing the next generation of tactical networking with the Department of Defense, Joint Tactical Radio System Program Executive Office. He retired as an Active Member of the Army Acquisition Corps. Upon retirement, he named as the Director of the United States Army International Technology Center-Pacific, from 2012 to 2018, and was responsible for technology search and coordinating cooperation opportunities in science and technology throughout the Pacific Region on behalf of the U.S. Army. He is currently the Division Chief of the Advanced Research Division, Counter WMD Technologies Department, Research and Development Directorate, Defense Threat Reduction Agency.

FRANK SKIRLO received the B.Sc. degree from the University of Florida, the M.Sc. degree in electrical engineering from the University of Colorado at Colorado Springs, the master’s degree in strategic studies from the U.S. Army War College, and the Ph.D. degree in electrical engineering from George Mason University, with the thesis focusing on image processing and UAS-based aided target detection. He is with the DTRA’s Advanced Research Division, Counter-WMD Department, Research and Development Directorate, as a Technical Program Analyst. He was a U.S. Army Signal Officer for over 24 years of active service, and retired at the rank of Colonel in December 2011, after completing a variety of worldwide assignments, including command and staff positions in Germany, South Korea, and Hawaii. His last assignment was with the joint staff J-6, overseeing C4 transport programs and requirements. After retiring from the military, he has supported the U.S. Army CSISR Centers Night Vision and Electronics Sensors (currently Research and Technology Integration) Directorate Integrated Sensor Architecture (ISA) Project.

* * *