IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 August 2023, accepted 19 August 2023, date of publication 25 August 2023, date of current version 30 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3308822

RESEARCH ARTICLE

Assessing the Predictive Performance of Two DNN Models: A Comparative Analysis to Support Reusing Training Weights for Autonomous Aerial Refueling Missions

VIOLET MWAFFO, DILLON MILLER, AND DONALD H. COSTELLO, III
Weapons, Robotics, and Control Engineering Department, United States Naval Academy, Annapolis, MD 21402, USA

Corresponding authors: Donald H. Costello, III (dcostell@usna.edu) and Violet Mwaffo (mwaffo@usna.edu)

This work was supported by the Department of the Navy, Office of Naval Research.

ABSTRACT The United States Navy aims to enhance its fleet by expanding the deployment of unmanned aircraft in carrier air wings. However, certifying the autonomous refueling of unmanned aerial platforms currently lacks a publicly available method. Ongoing research at the United States Naval Academy focuses on investigating certification evidence that would enable a deep neural network (DNN) to facilitate autonomous aerial refueling (AAR). This study explores training a DNN to accurately detect the drogue and coupler deployed by a KC-130 tanker and a tanker-configured F/A-18 jet. Both tankers have a similar drogue refueling system but differ vastly in image background noise and contrast, posing a challenge for object detection. Using salient metrics, the performance of a DNN model trained separately on video footage of both tankers is tested to enable the AAR task. Our results indicate that a DNN trained on developmental flight test videos of aircraft refueling from a KC-130 tanker effectively completes the aerial refueling task on a F/A-18 tanker compared to another DNN trained on video footage of the same tanker. These findings might validate the idea that a DNN trained on a specific aircraft dataset with a similar probe and drogue refueling system satisfactorily performs the aerial refueling task on various tankers, eliminating the need for additional training data for each tanker individually.

INDEX TERMS Autonomous aerial refueling (AAR), deep neural network (DNN), probe-and-drogue system, unmanned aerial vehicle (UAV).

I. INTRODUCTION

The United States Navy (USN) has made a public commitment to significantly augment the presence of unmanned platforms within deployed carrier air wings [1]. Consequently, naval authorities anticipate these aircraft will be capable of conducting aerial refueling operations. To address this objective, the United States Naval Academy (USNA), in collaboration with the Office of Naval Research (ONR) and the Naval Air System Command (NAVAIR), is actively engaged in research endeavors aimed at producing certification evidence. The ultimate goal is to obtain a safety of flight clearance that would authorize unmanned aircraft to autonomously perform aerial refueling tasks as the receiving aircraft.

The task of autonomous aerial refueling (AAR) presents significant challenges and currently lacks a viable solution for Unmanned Aerial Vehicles (AUVs) to function as the receiving aircraft. However, the ONR has provided funding for a potential pathway towards achieving AAR [2], incorporating relative navigation as described in References [3] and [4]. In this approach, the tanker aircraft would employ a data link to transmit precise location information regarding

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

the refueling drogue. The unmanned receiver would then utilize this data to maneuver into a predetermined pre-contact position positioned 5-25 feet directly behind the refueling drogue. Then a computer vision system would provide the necessary information to track the drogue and link the probe tip to the coupler.

Current research in vision-based AAR systems has investigated various techniques, such as incorporating light-emitting diodes (LEDs) and highly reflective materials, to enhance drogue detection [5], [6], [7]. However, these approaches require modifications to the drogue and rely on artificial features, introducing uncertainties and potential hazards during refueling. Additionally, they are susceptible to image loss in adverse weather conditions, turbulence, low visibility, and light interference.

Alternatively, non-artificial feature-based methods have been proposed. These include complex geometric procedures like template matching and threshold segmentation [8], monocular vision-based approaches using direct image registration [9], and techniques such as multiscale, low-rank, and sparse decomposition [10], [11]. However, these methods have primarily been tested under normal environmental conditions and may exhibit inaccuracies in complex situations involving clouds, fog, and light interference. Template matching and threshold segmentation, in particular, may require recalibration when environmental conditions deviate from the original calibration parameters.

Recent research efforts have explored the application of deep neural networks (DNNs), specifically convolutional neural networks (CNNs), for drogue identification and localization [12]. These innovative approaches leverage extensive databases of AAR aircraft to extract meaningful features for object detection. By employing DNNs, relevant features can be directly extracted from picture frames using state-of-the art CNN in order to directly classify objects without the need of using additional artifacts, geometric procedures, or to re-calibrate the model to comply to a given environmental condition. As a result, DNN-based AAR methods offer a more efficient and accurate drogue detection approach with high robustness and faster processing speed compared to classical methods that rely on artificial features or geometric procedures [12].

This study revolves around the utilization of a DNN to accurately detect the drogue and coupler deployed by the tanker aircraft. The underlying premise is that an UAV can employ a computer vision-based DNN to locate and maneuver its probe tip into the coupler from the pre-contact position. Therefore, a primary focus of this research is to train and validate the DNN's accuracy in accomplishing the task of aerial refueling. In addition, this work aims to assess the feasibility of utilizing a DNN trained on one aircraft system for a different system that employs a similar refueling procedure, without the need to retrain the DNN specifically for that system. As per the guidelines set forth by the ONR, the approach strictly avoids any modifications to the refueling drogue. Consequently, no additional elements such as infrared markers or special markings are employed to assist the computer vision-based DNN detection process.

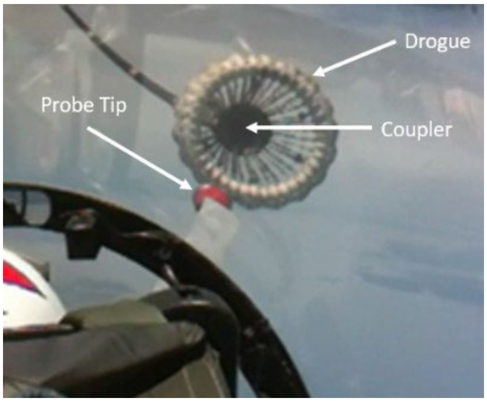

FIGURE 1. F/A-18F Super Hornet Preparing to Refuel [13].

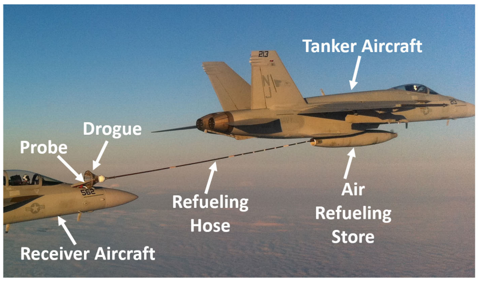

FIGURE 2. EA-18G Refueling from the F/A-18E Tanker [14].

Initial investigations conducted at the USNA have focused on employing a DNN to accurately detect the drogue and coupler located behind the wing refueling station of a KC-130 tanker aircraft. Figure 1 displays an image of a F/A-18 aircraft in preparation for aerial refueling from a KC-130, with clear identification of the refueling probe tip, drogue, and coupler. Figure 2 displays all of the relevant components required to enable probe and drogue aerial refueling. In Reference [15], preliminary works illustrating the possibility to utilize DNN training weights from the KC-130 tanker to detect the drogue and coupler behind a F/A-18 was documented. Building upon prior research, our current investigation aims to assess the performance of a DNN trained on the F/A-18 dataset in comparison to one trained on the KC-130 dataset when evaluated against a F/A-18 dataset. Specifically, we analyze the performance metrics of the DNNs trained respectively on the KC-130 and F/A-18 datasets, evaluating their weights against previously unseen F/A-18 datasets. The outcomes of this study are anticipated to yield significant time and resource savings when adapting existing resources to newer aircraft platforms. The main contributions of this paper are the following:

- segmenting the drogue into multiple classes and using only a few pre-processing procedures to resolve

background noise and low contrast compared to traditional AAR methods [12] in order to reduce single point failure and thus to improve the overall detection performance of the DNN without the need to consider several redundant and complex data augmentation procedures [16], [17], [18];
- demonstrating the possibility to extend a trained DNN model for AAR to a similar AAR object detection task on a different platform, allowing to significantly reduce the certification time and the high cost required for conducting additional experiments.

The structure of the paper is as follows: Section II provides a concise literature review and overview of object detection using DNNs. In Section III, we elaborate on the proposed investigative method, which includes details on the training dataset, model training, and evaluation procedure. Section IV presents the main findings of our study, highlighting the selected metrics used to assess the effectiveness of the trained DNN models. Finally, Section V concludes the paper by summarizing the work and suggesting potential directions for future research.

II. BACKGROUND ON VISION BASED OBJECT DETECTION

A brief overview of recent progress in computer vision using state-of-the-art DNN for object detection is provided in section II-A. Section II-B discusses data augmentation and Section II-C presents transfer learning.

A. OBJECT DETECTION IN COMPUTER VISION

In the field of computer vision, object detection [19] involves categorizing objects into predefined classes [20], [21] and estimating their spatial coordinates [22] within image frames. Recent advancements in DNNs have led to the availability of state-of-the-art object detection models [23], [24], [25], [26], significantly improving both accuracy and inference speed. These advancements have been facilitated by benchmark datasets such as ImageNet [24], PASCAL visual object classes (VOC) [25], and Microsoft common objects in context (MS-COCO) [26], as well as efficient backbone networks for feature extraction and powerful computing platforms. However, object detection for specialized tasks can present various challenges [27], including poor image quality, limited training data, variations in object scales [28], and the presence of cluttered or noisy backgrounds [27], [29], [30]. These challenges can lead to misclassifications or inconsistent classifications across video frames, which are not desirable for sensitive tasks such as AAR.

One popular single-stage object detection model is the so-called You Only Look Once (YOLO) [31]. YOLOv5 [32] is part of this family of object detectors and comprises three key components: a Backbone, Neck, and Head. The Backbone extracts multi-scale image features, the Neck combines these features to generate feature maps, and the Head performs class and bounding box predictions. YOLOv5 employs the CSPDarknet53 as its backbone network, which consists of 29 convolutional layers and a total of 27.6 million parameters. The particularity of the YOLOv5 backbone is the stacking of multiple CBS (Convolution + Batch Normalization + Sigmoid Linear Unit) and Concentrated-Comprehensive Convolution (C3) modules. The model also incorporates a Spatial Pyramid Pooling (SPP) block and Path aggregation network (PANet) for feature fusion allowing to enhance the richness of extracted features. At the last stage, YOLO implements a filter with a set of threshold values and the non-maximum suppression (nms) process to output the final detection information. When trained on benchmark datasets such as Pascal VOC [33] and MS-COCO [26], YOLOv5 achieves remarkable performance by leveraging innovative techniques such as self-adversarial training and cross-stage partial connections. There are five official versions of YOLOv5 depending of the network size or the number of parameters, namely extra-large YOLOv5x, large YOLOv5l, medium YOLOv5m, small YOLOv5s, and nano YOLOv5n.

B. DATA AUGMENTATION

Training a DNN necessitates a significant amount of data, which can often be challenging to gather notably to complete the AAR task [16]. In the literature, several techniques have been employed to generate novel or additional datasets. These procedures encompass various approaches, including data augmentation, synthetic data generation using specialized 3D computer graphics and physics engine software, and human-assisted laboratory experiments to replicate real-world scenarios.

Data augmentation involves creating synthetic datasets by applying affine or geometric transformations and employing additional computer vision techniques to introduce more variability into a limited real-world dataset [34]. When implemented effectively, this method enables the coverage of data distributions that may not be adequately represented in the original dataset, while maintaining consistency with real-world cases. Data augmentation has been demonstrated to significantly enhance detection models [35]. However, in practice, it may not be feasible to recreate a dataset that captures all variations observed in real-world images, as there are limitations to the amount of variability that can be added to real datasets. Consequently, when faced with limited and less diverse real data, data augmentation may lead to overfitting, where the model performs well on the training dataset but poorly on unseen data, or class imbalance, where the model tends to better predict the most represented classes while neglecting minority classes [36], [37].

C. TRANSFER LEARNING

Transfer learning is a powerful technique in machine learning and deep learning that involves leveraging knowledge acquired from pre-trained models to tackle new tasks or domains [38]. Rather than starting from scratch, transfer learning allows us to reuse the learned representations and knowledge of an existing model, which has been trained on a

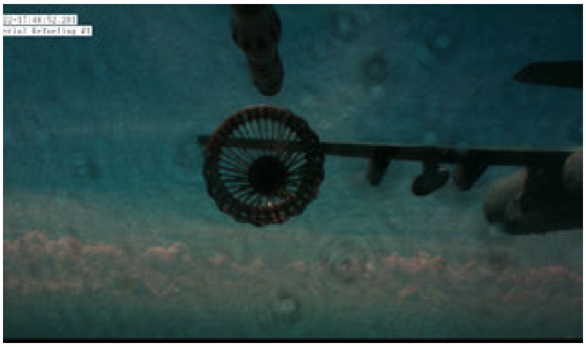

FIGURE 3. E-2D aircraft preparing to refuel off a KC-130 tanker [40].

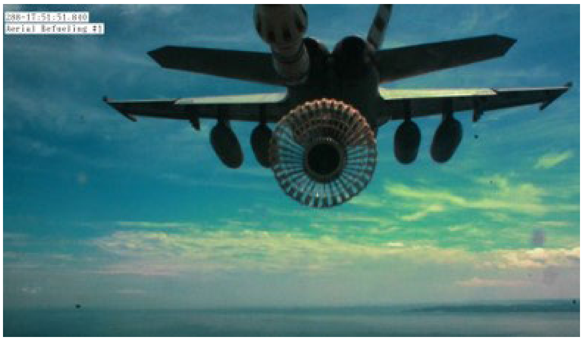

FIGURE 4. E-2D aircraft preparing to refuel off a F/A-18 tanker configured aircraft [40].

large-scale dataset like ImageNet [24], PASCAL VOC [33], or MS COCO [26]. By adapting and fine-tuning the pre-trained model on a smaller dataset specific to the new task, transfer learning enables faster and more efficient training while improving performance. A notable milestone in transfer learning is the residual network (ResNet) architecture [39], achieving state-of-the-art performance on the ImageNet dataset. In [39], transfer learning effectiveness was evidenced by fine-tuning pre-trained models for specific image classification tasks.

Transfer learning offers several advantages. It reduces the need for extensive labeled data and computational resources since the pre-trained model has already learned to extract general features from complex data, capturing valuable information that can be useful in the new task. By building upon these pre-existing representations, the model can quickly adapt and specialize for the unique nuances and characteristics of the new dataset. Transfer learning has gained wide popularity across various domains, enabling the application of deep learning to a broader range of problems and accelerating the development of new models and solutions.

III. DATASET, MODEL TRAINING, AND EVALUATION

This section provides a brief description of the data sets available to the USNA for this research effort, a review of the process used for the KC-130 trained DNN, and preliminary results about its performance against the F/A-18 data set.

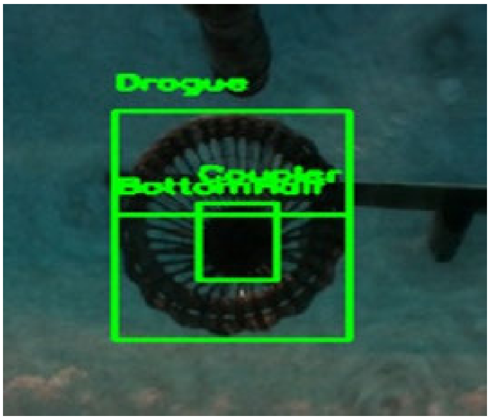

FIGURE 5. Sample KC-130 image from the labelled dataset [40].

A. E-2D DEVELOPMENTAL FLIGHT TEST VIDEO FOOTAGE

The given data set consists of 63 videos of aerial refueling engagements from the E-2D Hawkeye experimental test flights conducted by VX-20 at Naval Air Station Patuxent River, MD. These videos range from 1 to 5 minutes, and feature refueling conducted on two different aircraft tankers: the KC-130 and the F/A-18. The video footage are recorded at various time during daylight and weather conditions. Figure 3 is a screen capture from the KC-130 and Figure 4 is a screen capture from the F/A-18 aerial refueling video footage. In comparison to Figure 4, Figure 3 exhibits lower background noise and higher contrast where the drogue overlaps with a small portion of the wing deployed by the KC-130. Higher noise and lower contrast are observed in Figure 4 on the top half of the drogue which, in the picture, extends over to cover part of the air refueling store attached underneath the F/A-18. In addition, one can observe that the drogue utilized by either aircraft is a similar drogue cone shape with a coupler to link the probe tip and metallic struts to guide the probe tip but have different circular canopy and gore spacing. In addition, the drogue size is different between aircraft with a larger drogue size for the KC-130 tanker compared to the F/A-18 tanker. These contrasting features make it challenging to reuse a DNN object detector model trained on a particular aircraft system in order to extend it to another one.

B. MODEL TRAINING

We observed that the upper half portion of the drogue in the F/A-18 video footage displays noisy background and low contrast as it overlaps with the F/A-18 air refueling store making it difficult to detect the drogue. Figure 5 is a labeled image from the KC-130 data set and Figure 6 is a labeled image from the F/A-18 data set illustrating the issue observed. To resolve this problem, a traditional solution is to add some occlusion augmentation procedures to the training data in order to improve the DNN model's resilience. A few of these procedures include [16], [18] random erase where a random rectangular section of the image is erased and replaced with noisy pixels, cutout where random square image portions

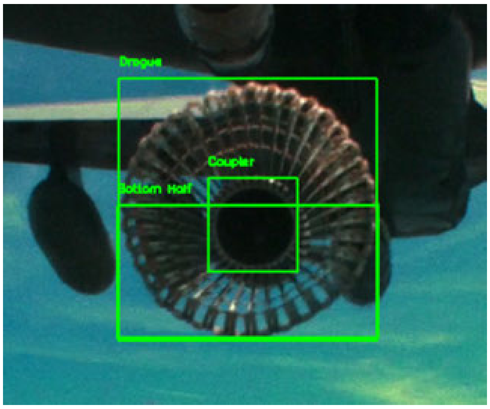

FIGURE 6. Sample F/A-18 image from the labelled dataset [40].

are removed from the image, hide-and-seek or grid-mask where a grid is drawn over the image and portions of the grid randomly hidden with some probability, cutmix where a portion of the image is randomly cut out and replaced with a portion of a different image. Another method is mosaic augmentation [17] where four different images are stitched together to form a picture which in turn is randomly shifted and cropped, allowing the model to learn to identify objects in different contexts and portions of the image. Some of these pre-processing procedures have been shown applicable for AAR tasks and notably to identify various aircraft type [16]. Others are automatically implemented in the YOLOv5 training pipeline. Our method is based on a drogue partition in three classes to be detected by the DNN including, the entire drogue, the lower half of the drogue, and the coupler. This partition is complemented by a few data augmentation procedures. This approach is easier to implement, prevent single point failure, and allow to avoid using any redundant procedures that might lead to over-fitting. We show that it provides good prediction results for extending a trained DNN weights to a different platform.

Prior to the training phase, a custom software was utilized to assist with data labeling (see Figure 7). The custom script initially down-samples the picture frames a rate of 1 out of 4 in each video file. This procedure allowed the removal of some very similar frames in order to prevent any issue related to over-representation of certain features in the training dataset as these might significantly bias the predictions of the object detection model. The custom-script relied on a semi-automated procedure where a picture frame was initially labeled by hand and then a few next N picture frames were automatically labeled where N was set to 20. This number was empirically determined by the operator after a few trials to ensure consistency and no significant movement of the drogue in the picture frames between steps. At the end of the labeling procedure, the operator has the possibility to manually review and correct the bounding boxes and any mislabelling cases. The images were finally augmented in terms of brightness (±25%), saturation (±25%), and scale

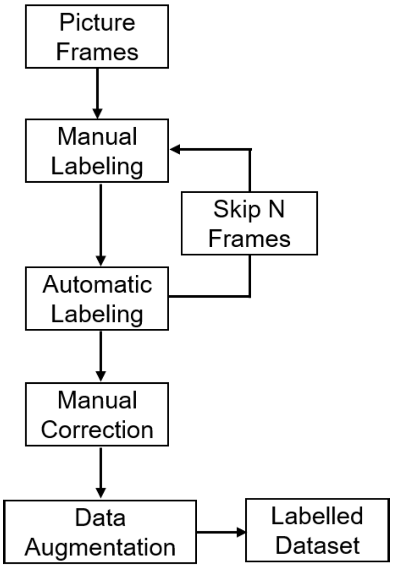

FIGURE 7. Schematic description of the custom labeling and data augmentation toolbox using an object detection model to infer bounding boxes on objects manually labeled in the next N picture frames.

(.3 to 2 randomly). These data augmentation procedures allowed us to increase the data set size and enhance the generalization ability of our object detection model with increased variability in the labeled data set images. The DNNs were trained to identify the three labeled classes including the entire drogue, the lower half of the drogue, and the coupler. The KC-130 and F/A-18 trained DNNs were the result of an iterative process to test the possibility of reusing the DNN training weight on a similar but different platform.

Model training relies on transfer learning using the state-of-the-art object detector YOLOv5. The small size model of YOLOv5 is retained for its reduced size (only 7.5 millions parameters) and its speed of execution while providing a quite good level of accuracy. Indeed, YOLOv5 backbone incorporates cross stage partial network (CSPNet) into Darknet [41] allowing to decrease the model parameters and floating-point operations per second (FLOPS). This integration not only allows to ensure greater inference speed and accuracy, but also significantly reduces the model size which is desirable when executed on limited edge computing devices. The PANet used as YOLOv5 neck [42] allows to boost the propagation of low-level features and to enhance small object localization. In addition, YOLOv5 neck generates 3 different sizes of feature maps in order to achieve multi-scale [43] prediction and for the model to handle small, medium, and large objects. This output structure of YOLOv5 neck is suitable to handle our partition of the drogue into three portions with different sizes. The total number of real and augmented images resulting from data preprocessing of the KC-130 video-footages was 2128 among which 1476 resulted from data augmented. The picture frames were subdivided into 1703 for training and 425 for validation. The DNNs were

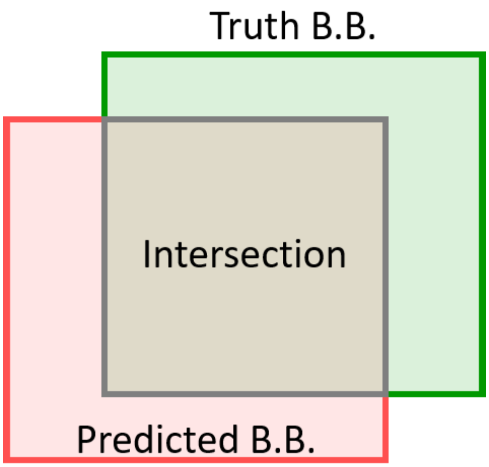

FIGURE 8. **Illustration of the IoU with the truth bounding box in green, the predicted bounding box in red, and the overlap or intersection area in gray.**

initially trained for a total of 200 and 1000 epochs. The latter was retained as it further improve object detection. Alternatively, hyperparameters tuning using a genetic algorithm such as the hyperparameters evolution method [44] could have been utilized to optimize model parameters. However, the later procedure was not implemented as it is complex and expensive due to the large search space and unknown dependencies between parameters.

The F/A-18 DNN was trained using a similar approach as the KC-130 trained DNN. The F/A-18 video footage were annotated with three classes including the whole drogue, the bottom half of the drogue, and the coupler. Data were augmented using a similar procedure in terms of brightness, saturation, and scale. This resulted into an identical number of augmented dataset which in turn was subdivided into the same proportion for training and validation.

C. MODELS EVALUATION

The performance of a DNN is often assessed based on the mean average precision (mAP) benchmark metric, widely used in the computer vision research community. Important metrics used when analyzing DNN performance using the mAP are intersection over union (IoU), precision, recall, and average precision (AP).

The IoU measures the overlap between the detected bounding box and the ground truth bounding box. It allows us to determine the accuracy of positive detection (see Figure 8) and is calculated using Equation (1) by dividing the intersection area, shared by the predicted bounding box and the true bounding box, with the union area, the total area covered by both the true bounding box and the predicted bounding box. This is represented in Equation (1), where A_{True} is the area of the true bounding box and $A_{Predicted}$ is the area of the predicted bounding box. The baseline acceptable threshold value for IoU is 0.5 [45], [46], considering a detection as "true positive" if the intersectional area is at least half of the ground true total area.

$$IoU = \frac{(A_{Predicted} \cap A_{True})}{A_{Predicted} + A_{True} - (A_{Predicted} \cap A_{True})} \tag{1}$$

Precision measures the accuracy of the model in correctly classifying objects, while recall determines the model's ability to identify objects when they appear in the image. Precision effectively quantifies the DNN's accuracy when making predictions and is calculated in Equation (2a) as the ratio of the total number of true positive detections by the sum of the true positive and false positive detections where the latter quantifies all irrelevant detections. Alternatively, the recall is a ratio allowing us to effectively quantifies the DNN's ability to predict an object that is known to exist from the truth data. Recall is calculated in Equation (2b) by dividing the total number of true positive detections by the sum of the true positive and false negative detections, effectively quantifying the DNN's ability to detect an object that is known to exist in a given frame.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{2a}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2b}$$

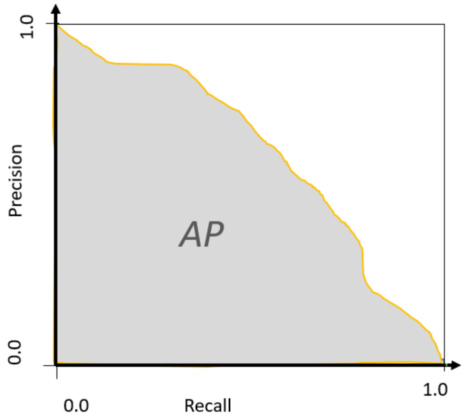

FIGURE 9. **Illustration of the Precision-Recall curve and Average Precision for a single class at a given IoU value.**

AP is calculated in Equation 3a for a single class as the area under the precision-recall curve (Figure 9) using a specific IoU threshold value, typically 0.5. To adequately quantify the performance of the DNN at all levels of bounding box accuracy, average precision is calculated across IoU threshold values ranging from 0.5 to 0.95 at increments of 0.05 where 0.5 is the minimum acceptable IoU value for object detection. The mAP combines precision and recall over all classes to provide a holistic measure of the DNN's performance. Two variants of the mAP are often quantified. The first quantity in Equation (3b) is denoted as mAP50 and indicates the mean average precision over all classes at IoU = 0.5 value. The second quantity is denoted as mAP50:95 to indicate the mAP over all classes at IoU between 0.5 to 0.95. The former is often used as the minimum baseline threshold value [45] for object detection while the latter provides an evaluation of the overall performance of the model at all IoU threshold values [46].

$$AP = \int_0^1 p(r)\,dr \tag{3a}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{3b}$$

The F1-score [45] is defined as the harmonic mean of the precision and recall of an object detector and determined for each class i by the expression:

$$\mathrm{F}_{1i} = 2\frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

The F1-score values are in the interval [0, 1] with values closer to 0 corresponding to worst cases where precision or recall are closer to 0 and values closer to 1 when both precision and recall values are closer to 1. For a multi-class dataset, the macro-averaged value of the harmonic mean of the class-wise precision and recall values can be computed for dataset with an identical number of classes as $\mathrm{F}_1 = \frac{1}{N}\sum_{i=1}^{N}\mathrm{F}_{1i}$. The confidence threshold value maximizing the F_1-score allows to determine the ideal object detector model with the highest precision and recall. In turn, combined with the selected IoU threshold value, it can be utilized in production for real world implementation of the trained model.

Finally, to evaluate the relevance of reusing the AAR DNN weights accross platforms, the trained DNN models which are (i) YOLOv5 trained on the KC-130 video footage (KC-130 trained DNN) and (ii) YOLOv5 trained on the F/A-18 video footage (F/A-18 trained DNN), were tested on unseen developmental flight test footage of the E-2D attempting to aerial refuel behind a F/A-18 aircraft. The metrics retained to evaluate the performance of the trained DNNs are the precision, recall, the mAP50 evaluated at the IoU value of 0.5, the more general mAP50:95, and the F1-score. The test is said to be successful if both DNNs taken separately perform well and in addition, the KC-130 trained DNN performs as well as the F/A-18 trained DNN when tested on the same unseen F/A-18 dataset. In other words, the metrics for the F/A-18 trained DNN are the control values and those of the KC-130 trained DNN are the test values.

IV. MAIN RESULTS

Our results are summarized in Table 1 and Table 2, which present the precision, recall, mAP50, and mAP50:95 metrics for the KC-130 and F/A-18 trained DNNs over 1000 epochs.

TABLE 1. Training evaluation results of the KC-130 and the F/A-18 trained DNNs using 1000 epochs.

All class	P	R	mAP50	mAP50:95
KC-130 DNN	0.99965	0.98824	0.98982	0.91964
F/A-18 DNN	0.99925	0.99421	0.99225	0.91927

TABLE 2. Prediction test results with the KC-130 and the F/A-18 trained DNNs on unseen video footage of the F/A-18 refueling aircraft measured by the Precision, Recall, and mAP.

KC-130 DNN	IoU = 0.5			IoU∈[0.5:0.95]
Class	P	R	mAP50	mAP50:95
all	0.996	0.907	0.977	0.634
Coupler	0.995	0.984	0.987	0.512
BottomHalf	1	0.871	0.978	0.722
Drogue	0.994	0.866	0.965	0.668
F/A-18 DNN	IoU = 0.5			IoU∈[0.5:0.95]
Class	P	R	mAP50	mAP50:95
all	0.999	0.994	0.994	0.787
Coupler	0.996	0.99	0.992	0.6
BottomHalf	1	0.998	0.995	0.826
Drogue	1	0.994	0.995	0.935

Table 1 showcases the performance metrics immediately after training, evaluating the models on the same aircraft models they were trained on. The results indicate that both trained DNNs perform well on the respective tankers they were trained on. They achieve precision and recall rates close to 99%, a mAP50 value also close to 99% for both models and a more comprehensive mAP50:95 value close to 92%.

Table 2 presents the performance metrics obtained by testing both trained models on unseen video footage of an E-2D aircraft refueling behind a F/A-18 tanker. A total of 195 unseen picture frames were utilized for testing the trained DNN models with the raw data displaying a cloudy and less bright sky typical to late afternoon times as shown in Figure 2(b). Prior to the test, the picture frames were labelled following a similar procedure as for the trained and validation datasets. Here, both trained DNN models exhibit a high precision rate close to 99%. They also achieve high mAP50 values, with the F/A-18 trained DNN achieving a mAP50 value of 99.5% and the KC-130 trained DNN achieving a mAP50 value of approximately 98%. The slight difference in the mAP50 value observed here is primarily due to a lower recall value for the KC-130 trained DNN compared to the F/A-18 trained DNN. While both DNNs demonstrate overall good recall values, the F/A-18 trained DNN is without any doubt more effective on F/A-18 video footage, with an overall recall value close to 99%, while the KC-130 trained DNN achieves a slightly lower recall value of 91%. This suggests a higher rate of false negatives among the total predictions for the KC-130 trained DNN. Consequently, the more comprehensive mAP50:95 is estimated at around 63.4% for the KC-130 trained DNN, while it achieves a much higher value at 79% for the F/A-18 trained DNN.

Both DNNs demonstrate good accuracy in predicting all three classes: the entire drogue, the bottom half, and the coupler. When analyzing individual classes using the trained F/A-18 DNN model, we observe that, when evaluated on the same aircraft model it was trained on, the entire drogue is more accurately predicted compared to the bottom half drogue and the coupler. This is because the drogue, being larger and providing more distinctive features, results in higher prediction accuracy. However, when evaluating the

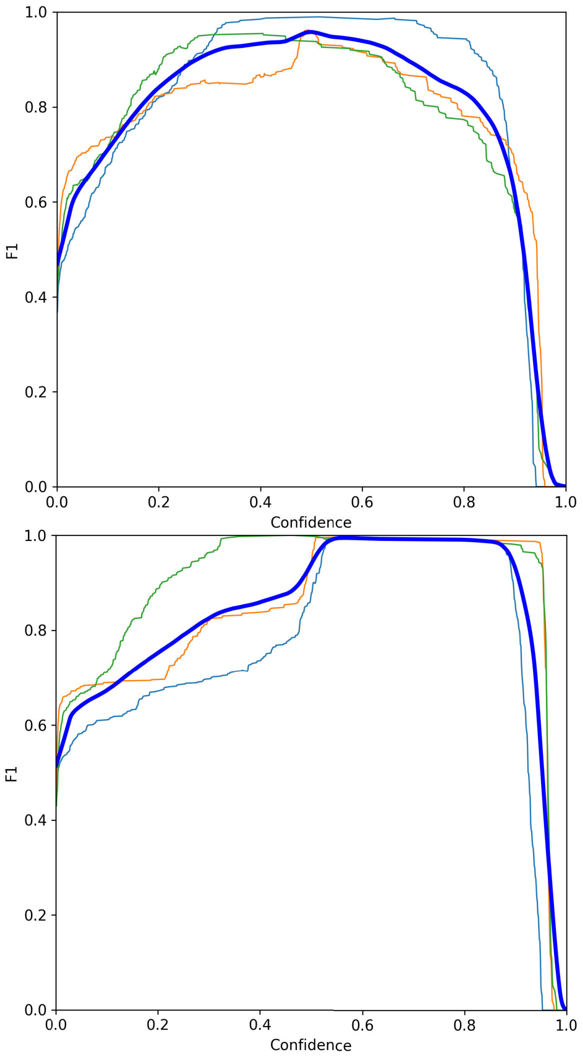

FIGURE 10. **F1 confidence curve of the KC-130 (top) and the F/A-18 (bottom) trained DNN tested on the F/A-18 aerial refueling tanker picture frames. The curve in light blue is the curve for the coupler, orange is for the bottom half drogue, green is for the whole drogue, and dark blue is all classes combined.**

KC-130 trained DNN on video footage of the F/A-18 tanker, which is a completely different aircraft model, our results demonstrate that the bottom half drogue is more accurately predicted, with a comprehensive mAP50:95 value estimated at 72.2%, compared to the accuracy value of 66.8% observed for the entire drogue.

At the minimum IoU of 0.5, the F/A-18 trained DNN shows high precision and recall values, while the KC-130 trained DNN exhibits high precision values for all classes but lower recall values for both the entire drogue and the bottom half drogue at about 87%. Therefore, the KC-130 trained DNN tends to confuse some elements of the entire drogue and the bottom half drogue with the background. Overall, the findings from this research demonstrates that the KC-130 trained DNN displays comparable performance to the F/A-18 trained DNN model in predicting the three classes, with some variations in accuracy depending on the aircraft type and specific part of the drogue being evaluated.

Further analysis were conducted using the F1 confidence curve in order to identify an ideal confidence threshold value that might be utilized to implement the object detector model. Figure 10 depicts various trends for the F1 confidence curves as the confidence level varies notably for individual F1 curve compared to the overall curve in dark blue. The individual F1 curves show that for the F/A-18 trained DNN, the whole drogue tend to be better predicted at a much more wider range of confidence level while for the KC-130 trained DNN it is the coupler which tend to display such performance. Such behavior is not surprising as the coupler for both aircraft tanker displays identical features compared to other parts of the drogue. In addition, it allows to improve the accuracy of the model at larger confidence levels which is relevant for reusing the KC-130 trained DNN model on AAR on the F/A-18 tanker. Our calculations also show that the overall F1 curve for all three classes achieves a peak value of 0.96 at a confidence level closer to 0.50 for the KC-130 trained DNN model while it displays a plateau closer to 1 for confidence values from 0.57 to 0.80 for the F/A-18 trained DNN model.

V. CONCLUSION

This work conducted a comparative analysis of the performance between a DNN trained on KC-130 data and another one trained on F/A-18 data for identifying the aerial refueling drogue and coupler behind a F/A-18. The evaluation of the trained DNN models, KC-130 and F/A-18, utilized precision, recall, mAP50, and mAP50:95 metrics. These metrics were complemented by a F1 confidence curve allowing to determine an optimum confidence level maximizing both precision and recall. When tested on their respective trained aircraft models, both models demonstrated excellent performance with precision and recall rates closer to 99%, along with high mAP values. However, when tested on unseen video footage of an E-2D refueling behind a F/A-18 tanker, the KC-130 model exhibited slightly lower recall and mAP50:95 values compared to the F/A-18 model, indicating a higher false negative rate. Our analysis revealed that the F/A-18 model outperformed the KC-130 model in predicting the larger drogue class. In contrast, the KC-130 model excelled in predicting the smaller bottom half drogue and coupler classes, as the upper drogue in KC-130 data is less noisy and has a better contrast. Both models displayed high precision for all classes, suggesting accurate detection with a negligible false positive rate when the desired object is present in the image. The ideal confidence level for reusing the KC-130 trained DNN to make predictions on the F/A-18 video footage was finally identified at a value closer to 0.5. The results underscored however the effectiveness of both DNN models in class prediction, with variations depending on the aircraft models and specific parts of the drogue being evaluated. Future works might explore additional data preprocessing procedures [16] to decrease the background noise and increase the contrast on AAR datasets. This includes the possibility to utilize appropriate filtering and augmentation procedures, drogue partition, image segmentation, infrared cameras for nightly or darker environments, or to complement object detection based DNN with a tracking algorithm in case of failure. Further research will also focus on optimizing drogue segmentation to enhance the possibility to reuse DNN training weights and extend the applicability of DNN models trained on US Navy aircraft to other NATO aircraft tankers using the probe and drogue refueling system.

ACKNOWLEDGMENT

The authors are thankful to Prof. R. Broussard for providing the semi-automated data labeling and augmentation toolbox.

REFERENCES

[1] P. Tucker. (Mar. 2021). *Drones Could One Day Make Up 40 Percent of a Carrier Air Wing, Navy Says*. [Online]. Available: https://www.defenseone.com/technology/2021/03/drones-could-one-day-make-40-carrier-air-wing-navy-says/172799/

[2] D. Costello, "Towards autonomous aerial refueling," in *Proc. Aerial Refueling Syst. Advisory Group Int. Meeting*, Orlando, FL, USA, 2023.

[3] J. Parry and S. Hubbard, "Review of sensor technology to support automated air-to-air refueling of a probe configured uncrewed aircraft," *Sensors*, vol. 23, no. 2, p. 995, Jan. 2023.

[4] D. Costello and H. Xu, "Using a run time assurance approach for certifying autonomy within naval aviation," *Syst. Eng.*, vol. 26, no. 3, pp. 271–278, May 2023.

[5] W. Xufeng, D. Xinmin, and K. Xingwei, "Feature recognition and tracking of aircraft tanker and refueling drogue for UAV aerial refueling," in *Proc. 25th Chin. Control Decis. Conf. (CCDC)*, May 2013, pp. 2057–2062.

[6] X. Wang, X. Kong, J. Zhi, Y. Chen, and X. Dong, "Real-time drogue recognition and 3D locating for UAV autonomous aerial refueling based on monocular machine vision," *Chin. J. Aeronaut.*, vol. 28, no. 6, pp. 1667–1675, Dec. 2015.

[7] C.-I. Chen, R. Koseluk, C. Buchanan, A. Duerner, B. Jeppesen, and H. Laux, "Autonomous aerial refueling ground test demonstration—A sensor-in-the-loop, non-tracking method," *Sensors*, vol. 15, no. 5, pp. 10948–10972, May 2015.

[8] Y. Yin, D. Xu, X. Wang, and M. Bai, "Detection and tracking strategies for autonomous aerial refuelling tasks based on monocular vision," *Int. J. Adv. Robotic Syst.*, vol. 11, no. 7, p. 97, Jul. 2014.

[9] C. Martínez, T. Richardson, P. Thomas, J. L. D. Bois, and P. Campoy, "A vision-based strategy for autonomous aerial refueling tasks," *Robot. Auto. Syst.*, vol. 61, no. 8, pp. 876–895, Aug. 2013.

[10] S. Chunhua, G. Shibo, and C. Yongmei, "Drogue detection algorithm in visual navigation system for autonomous aerial refueling," *Infr. Laser Eng.*, vol. 42, no. 4, pp. 1089–1094, 2013.

[11] S. Gao, Y. Cheng, and C. Song, "Drogue detection for vision-based autonomous aerial refueling via low rank and sparse decomposition with multiple features," *Infr. Phys. Technol.*, vol. 60, pp. 266–274, Sep. 2013.

[12] X. Wang, X. Dong, X. Kong, J. Li, and B. Zhang, "Drogue detection for autonomous aerial refueling based on convolutional neural networks," *Chin. J. Aeronaut.*, vol. 30, no. 1, pp. 380–390, Feb. 2017.

[13] D. Costello, *F/A-18F Preparing to Aerial Refuel Over Maryland in 2010. Several Key Elements for Refueling are Identified*, From the private collection of CDR Costello, 2021.

[14] *EA-18G Refueling From a F/A-18E*, From the private collection of CDR Costello, 2011.

[15] D. H. Costello, M. Violet, and D. Miller, "Transference training for a DNN to complete aerial refueling task," in *Proc. AIAA SCITECH*, Jan. 2023, p. 192, doi: 10.2514/6.2023-0192.

[16] R. Mash, B. Borghetti, and J. Pecarina, "Improved aircraft recognition for aerial refueling through data augmentation in convolutional neural networks," in *Proc. Int. Symp. Vis. Comput.*, Las Vegas, NV, USA, 2016, pp. 113–122.

[17] P. Kaur, B. S. Khehra, and Er. B. S. Mavi, "Data augmentation for object detection: A review," in *Proc. IEEE Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2021, pp. 537–543.

[18] M. Xu, S. Yoon, A. Fuentes, and D. S. Park, "A comprehensive survey of image augmentation techniques for deep learning," *Pattern Recognit.*, vol. 137, May 2023, Art. no. 109347.

[19] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.

[20] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017.

[21] J. Wang and L. Perez, "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Netw. Vis. Recognit.*, vol. 11, pp. 1–8, Dec. 2017.

[22] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 648–656.

[23] S. Vicente, J. Carreira, L. Agapito, and J. Batista, "Reconstructing Pascal VOC," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 41–48.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[25] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[27] Z. Li, Y. Wang, N. Zhang, Y. Zhang, Z. Zhao, D. Xu, G. Ben, and Y. Gao, "Deep learning-based object detection techniques for remote sensing images: A survey," *Remote Sens.*, vol. 14, no. 10, p. 2385, May 2022.

[28] D. Yu and S. Ji, "A new spatial-oriented object detection framework for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4407416.

[29] W. Ma, N. Li, H. Zhu, L. Jiao, X. Tang, Y. Guo, and B. Hou, "Feature split–merge–enhancement network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5616217.

[30] Y. Liu, Q. Li, Y. Yuan, Q. Du, and Q. Wang, "ABNet: Adaptive balanced network for multiscale object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5614914.

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

[32] G. Jocher, "YOLOv5 by ultralytics (Version 7.0)," Tech. Rep., 2020, doi: 10.5281/zenodo.3908559.

[33] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[34] K. Payumo, A. Huyen, L. Seguin, T. T. Lu, E. Chow, and G. Torres, "Augmented reality data generation for training deep learning neural network," *Proc. SPIE*, vol. 10649, pp. 232–243, 2018.

[35] X. Wang, K. Wang, and S. Lian, "A survey on face data augmentation for the training of deep neural networks," *Neural Comput. Appl.*, vol. 32, no. 19, pp. 15503–15531, Oct. 2020.

[36] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2018.

[37] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1–54, Dec. 2019.

[38] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Imag.*, vol. 9, no. 4, pp. 611–629, Aug. 2018.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[40] M. Machnicki, "NavAir public release 2022–601," NavAir, Patuxent River, MD, USA, Tech. Rep., 2022.

[41] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1571–1580.

[42] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "PANet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9196–9205.

[43] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

[44] G. Dai, L. Hu, J. Fan, S. Yan, and R. Li, "A deep learning-based object detection scheme by improving YOLOv5 for sprouted potatoes datasets," *IEEE Access*, vol. 10, pp. 85416–85428, 2022.

[45] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. D. Silva, "A comparative analysis of object detection metrics with a companion open-source toolkit," *Electronics*, vol. 10, no. 3, p. 279, Jan. 2021.

[46] 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.

VIOLET MWAFFO received the Ph.D. degree in mechanical engineering from the School of Engineering, New York University, in 2017. He is currently an Assistant Professor with the Weapons, Robotics, and Control Engineering Department, United States Naval Academic. His Ph.D. research, supported by the National Science Foundation, has focused on data-driven modeling and analysis of collective behavior observed in biological groups. His postdoctoral preparation, supported by the Chancellor Fellowship of the University of Colorado, has explored recent advances in theoretical and experimental study of multi-robotic systems. His research interests include bio-inspired systems, autonomous distributed systems, and artificial intelligence. He was a recipient of the National Science Foundation GK-12 Fellowship in Applying Mechatronics to Promote Science and Engineering, in 2012; an MITSUI-USA Foundation Fellowship, in 2015; and the Chancellor Postdoctoral Fellowship of the University of Colorado Boulder, in 2017.

DILLON MILLER was born in Lusby, MD, USA, in 2002. He is currently pursuing the bachelor's degree (Hons.) with the Weapons, Robotics, and Control Engineering Department, United States Naval Academy. He will pursue the M.S. degree in robotics engineering with the University of Maryland, College Park, MD, USA, as part of the voluntary graduate education program (VGEP). He is involved in the training and use of computer vision for application on UAVs.

DONALD H. COSTELLO, III, was born in San Diego, CA, USA, in 1978. He received the B.S. degree in systems engineering from the United States Naval Academy, Annapolis, MD, USA, in 2000, the M.A.S. degree in aeronautical science from Embry–Riddle Aeronautical University, Daytona Beach, FL, USA in 2005, the M.S. degree in aeronautical engineering from the Air Force Institute of Technology, Dayton, OH, USA, in 2009, the M.S. degree in systems engineering from the Naval Postgraduate School, Monterey, CA, USA, in 2011, and the Ph.D. degree in mechanical engineering from the University of Maryland, College Park, MD, USA, in 2020. He is currently an Assistant/Permanent Military Professor with the Weapons, Robotics, and Control Engineering Department, United States Naval Academy. He is involved in the certification and development of unmanned autonomous systems for practical use.