AIAA SciTech Forum
11–15 &amp; 19–21 January 2021, VIRTUAL EVENT
AIAA Scitech 2021 Forum

10.2514/6.2021-0397

Real-time Human Detection with Integration of Visual and Thermal Data from High Altitude sUAS

Chris Hayner*, Timothy Zhou†, Neil Gupta‡, Echo Liu§, Parker Mayhew¶ and Juris Vagners‖
Autonomous Flight Systems Laboratory, University of Washington, Seattle, WA, 98195, United States

A major challenge with Search and Rescue operations is covering a large amount of area quickly, efficiently, and accurately. With the rising prevalence of both computer vision algorithms for object detection and small Unmanned Aerial Systems (sUAS), an effective solution emerges. This work proposes a methodology for the integration of a sUAS, sensors, and software into a search and rescue workflow. Compared to ground search and rescue operations, sUAS allow for larger and faster area coverage at a fraction of the cost of manned aircraft missions. When combined with a versatile suite of object detection algorithms incorporating aerial data from both the visual and radiometric spectrum, the proposed system is able to analyze and process real-time data faster than existing systems at greater or comparable accuracy.

I. Introduction

A. Problem Statement

Search and rescue operation expenses are overwhelmingly dominated by personnel costs—a two-day Wilderness Search and Rescue (WiSAR) operation by the Chelan County Sheriffs Office (CCSO) in Enchantments, Washington saw $8,765.30 in personnel costs from just a $10,717.90 total (Section VII.D). Due to large areas and potentially rugged terrain, camera-equipped small unmanned aerial systems (sUAS) have emerged as a cheap and robust alternatives [1]. Able to rapidly provide near-ground imagery, sUAS operate several times faster than conventional methods—however, existing workflows involving manual extraction of visual information by human overseers are fallible to operator error and cognitive fatigue [2, 3]. As a result, such approaches often rely on redundancy provided by the use of multiple operators, heavily inflating expenditures [4]. In this paper, we suggest that a machine-learning based approach offers an effective solution, and introduce novel workflows for integrating a camera-equipped sUAS with Computer Vision (CV) algorithms for analyzing live aerial imagery. In particular, we will demonstrate:

1) the effective application of object detection frameworks (e.g. [5]) to aerial imagery using our custom-built dataset
2) successful integration of radiometric data streams and image processing algorithms to enhance detection performance
3) adaptability to a wide variety of wilderness environments and conditions

II. Literature Review

A. Previous Work at the University of Washington

Previous work done at the Autonomous Flight Systems Laboratory (AFSL) has focused on developing a variety of sUAS integrated systems, focusing in particular on aerial mapping [6, 7] and surveying [8, 9]. Of specific interest is the exhaustive search algorithm presented by Lum, Vagners, and Rysdyk [10, 11] which has potential applications in WiSAR sUAS as a basis for intelligent path planning and a scalable multi-agent architecture.

[*] Undergraduate Researcher, Department of Physics, University of Washington, Seattle, WA 98195, USA, AIAA Student Member
[†] Undergraduate Researcher, Department of Computer Science, University of Illinois at Urbana–Champaign, Champaign, IL 61820, USA
[‡] Undergraduate Researcher, College of Information & Computer Sciences, University of Massachusetts Amherst, Amherst, MA 01003, USA
[§] Graduate Researcher, Department of Applied Mathematics, University of Washington, Seattle, WA 98195, USA
[¶] Undergraduate Researcher, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[‖] Professor Emeritus, Department of Aeronautics and Astronautics, University of Washington, Seattle, WA 98195, USA, AIAA Member

B. Related Works

Title 14, Aeronautics and Space §107.19 [12] requires a Pilot in Command (PIC) to be present for non-recreational operation of sUAS in the United States. As the primary purpose of the PIC is the monitoring of sUAS operations, a secondary operator is often used to monitor payloads that require constant observation, such as live video feeds from Search and Rescue (SAR) payloads [13]. An overview by Adams et al. addressed how human operators interact with sUAS and how sUAS can be integrated into the WiSAR workflow. The overview further identified multiple models of differing implementations of the sUAS to show what an sUAS enabled WiSAR operation would look like. A 2008 study by Goodrich et al. demonstrated this methodology. A fixed-wing sUAS equipped with a video camera is used to address "the need to effectively present video information" and "the need to coordinate the UAV with ground searchers" in WiSAR operations. A CV algorithm described in [15] was used to extract features in real time.

With the rise of sUAS popularity in SAR operations, payload options have also seen an increase in performance and decrease in size. Thermal sensors have proven to be effective at identifying humans [16] but are limited in their resolution when compared to their visual counter-parts. A study done by de Oliveira and Wehrmeister in 2018 [17] focused on the use of low-cost equipment, specifically a Raspicam with a resolution of 2592 × 1944 pixels and a FLIR Lepton Long Wave Infrared sensor with a resolution of 80 × 60 pixels, for human identification via deep learning. For WiSAR applications, high resolution visual and thermal sensors are advantageous in that they allows sUAS to fly at higher altitudes, therefore getting more area in frame while maintaining the same number of pixels per human [1]. Multiple manufacturers such as FLIR and Workswell now offer options catered specifically to sUAS integrated SAR operations that have both a visual and thermal sensors in the same unit [18, 19].

Object detection is a main area of CV with diverse applications. For WiSAR, we are trying to detect human subjects to aid in rescue tasks. Within the detection procedure, feature extraction and classifiers play important roles, and selecting the right features gives more efficient and accurate detection, since features usually encode knowledge on objects, which are difficult to learn from a raw finite set of input pixels [20]. Wavelets, such as those proposed by Haar, have been shown to be very useful in extracting features [20]. Similarly, Convolutional Neural Networks (CNNs) are also used to extract features from images as seen in [21]. With the development of CNNs, such high-dimensional features can be extracted and reduced to low-dimensional features [22]. However, training a CNN with a large number of parameters requires a large dataset [23] to optimize the process. Advances in GPU hardware make it possible for such models relying on CNN architecture to perform well [24]. Numerous research efforts have shown promising results of human detection using visual or thermal image input from a sUAS view [25–27]. To use more information, researchers also incorporate thermal and visual inputs [17, 28, 29]. Our research efforts are similarly directed towards combining thermal and visual information to increase detection accuracy and robustness to non-ideal environments.

III. UW WiSAR Dataset

A. Dataset Acquisition

The UW WiSAR dataset includes over 30,000 visual and 5,500 thermal images taken by aerial sUAS over wilderness environments containing human subjects. The dataset was created with an emphasis on encompassing a wide variety of terrain, times of day, angles, and altitudes ranging from 6 m to 120 m. Example images from the dataset are shown in Fig. 1.

To prepare the dataset, human subjects within collected images were annotated using the YoloMark software [30]. While largely done by hand, the tool's object tracking features were utilized to reduce the time and manpower necessary to annotate the large dataset. To supplement our data, we used 770 of the more rural images from the VisDrone dataset [31, 32]. These include upwards of 15 human subjects per image whereas our images only had between 1 and 4 human subjects. One of the limitations of this dataset is that we rarely encountered wild animals that we could use as true negative examples during training, which would help to avoid potentially identifying wildlife as humans. In order to compensate, we explicitly added examples of humans walking with dogs.

B. Dataset Training

1. Pre-Training

For training, we used the method described in [33]. Because our algorithms also run detections on tiled images during inference, it is beneficial to train our models on tiles of the same size. As a result, visual images were split into

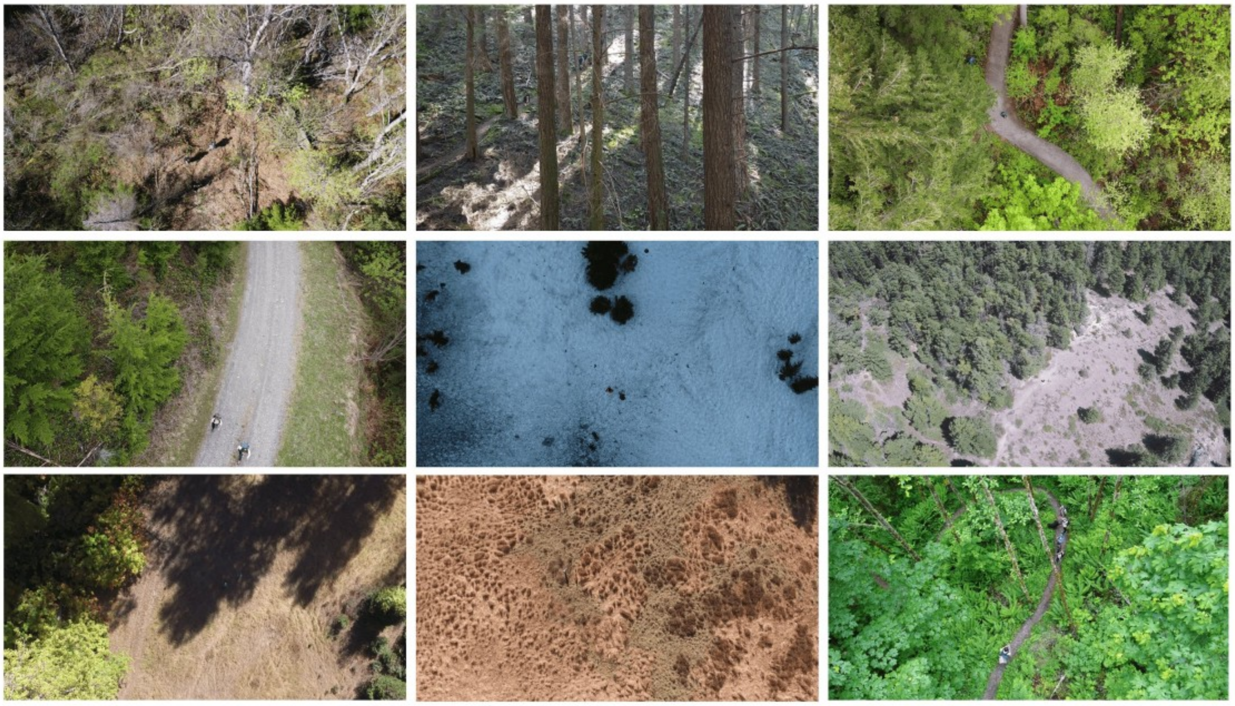


Fig. 1 Images from UW WiSAR dataset.

Table 1 Number of images in UW WiSAR Dataset

Dataset	Pre-Culled	Training set	Test set
Visual	211,228	27,435	3,026
Thermal	8,896	5,255	564

512 × 512 tiles (see Fig. 2). As the majority of images in our dataset contain only a small number of human subjects, the vast majority of these tiles were devoid of detections—roughly 85% of these non-salient tiles were culled from the dataset to reduce training time. By contrast, because radiometric data is already sufficiently low resolution 640 × 512 for direct inference, the model was trained on without further processing. For both thermal and visual data, the dataset was then split in approximately a 9:1 ratio, with the majority going into the training set and the remaining minority going into the test set as seen in Table 1.

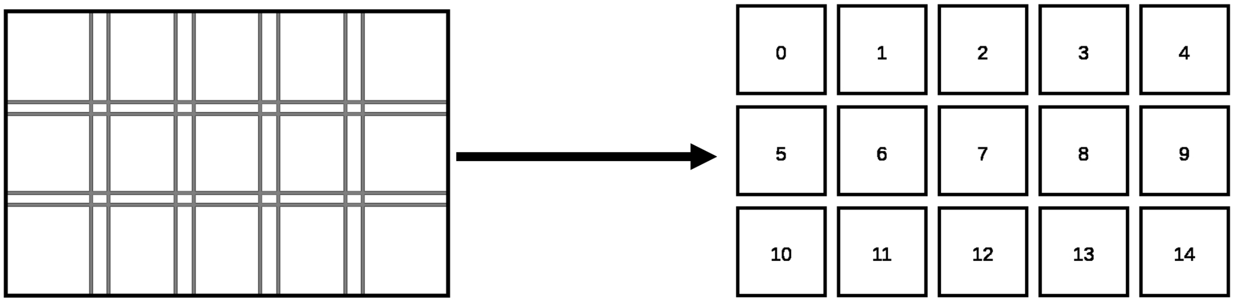


Fig. 2 Tiling Method.

2. Dataset Augmentation

In order to increase the variability of the dataset, YOLOv5 performs augmentation on the training and test sets, varying hue, saturation, value, rotation, translation, scale, and other parameters to decrease propensity for over-fitting. The batch shown in Fig. 3 demonstrates these augmentations. Bochkovskiy, Wang, and Liao detail this process more thoroughly in [34].

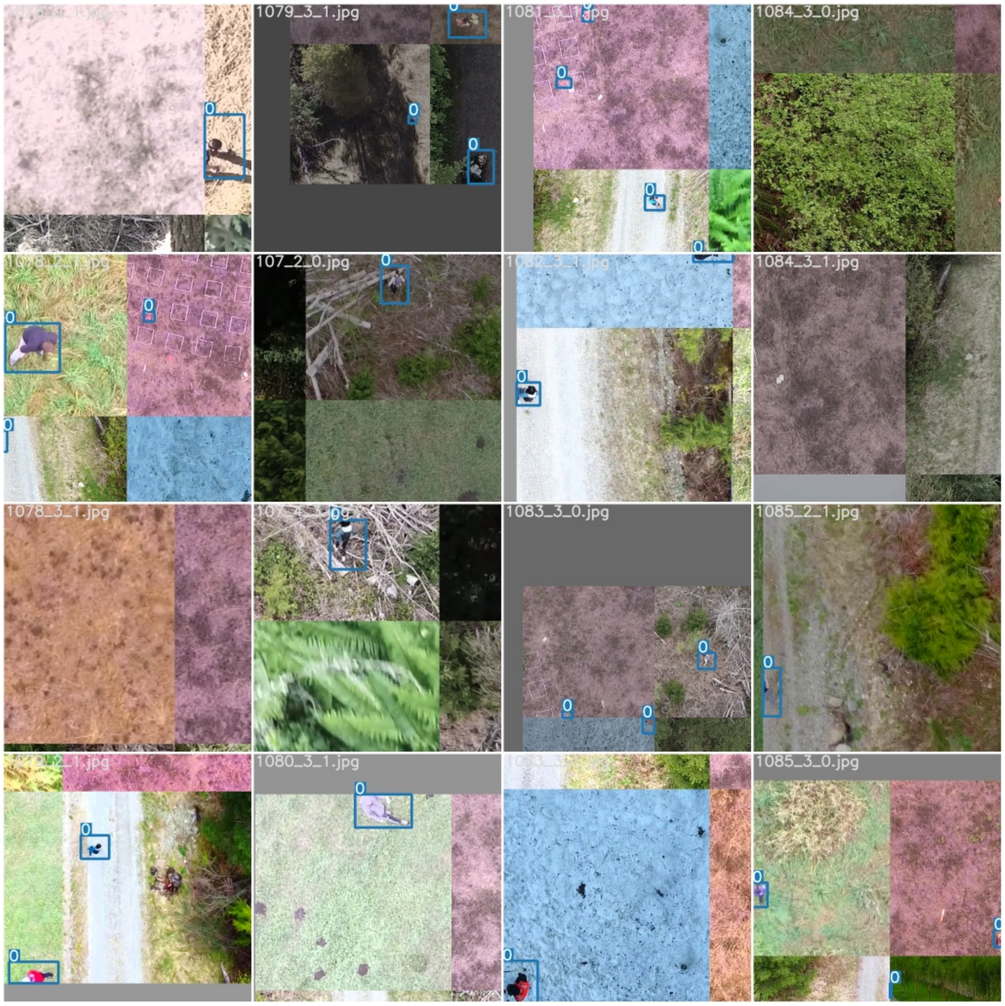


Fig. 3 Annotated Training Batch from UW WiSAR dataset.

IV. System Architecture

A. Overall System Architecture

The system is comprised of two main subsystems: an aerial unit and a ground unit. The aerial unit is an sUAS housing a radiometric and visual sensor package as well as a 5.8 GHz video transmission system providing live imagery to the ground unit. The ground unit is comprised of three main elements: the Remote Pilot in Command, an emergency response team, and a data processing unit which analyzes incoming data. In the event of a positive detection, Global Navigation Satellite System (GNSS) data is provided to the emergency response team from the aerial unit for further action.

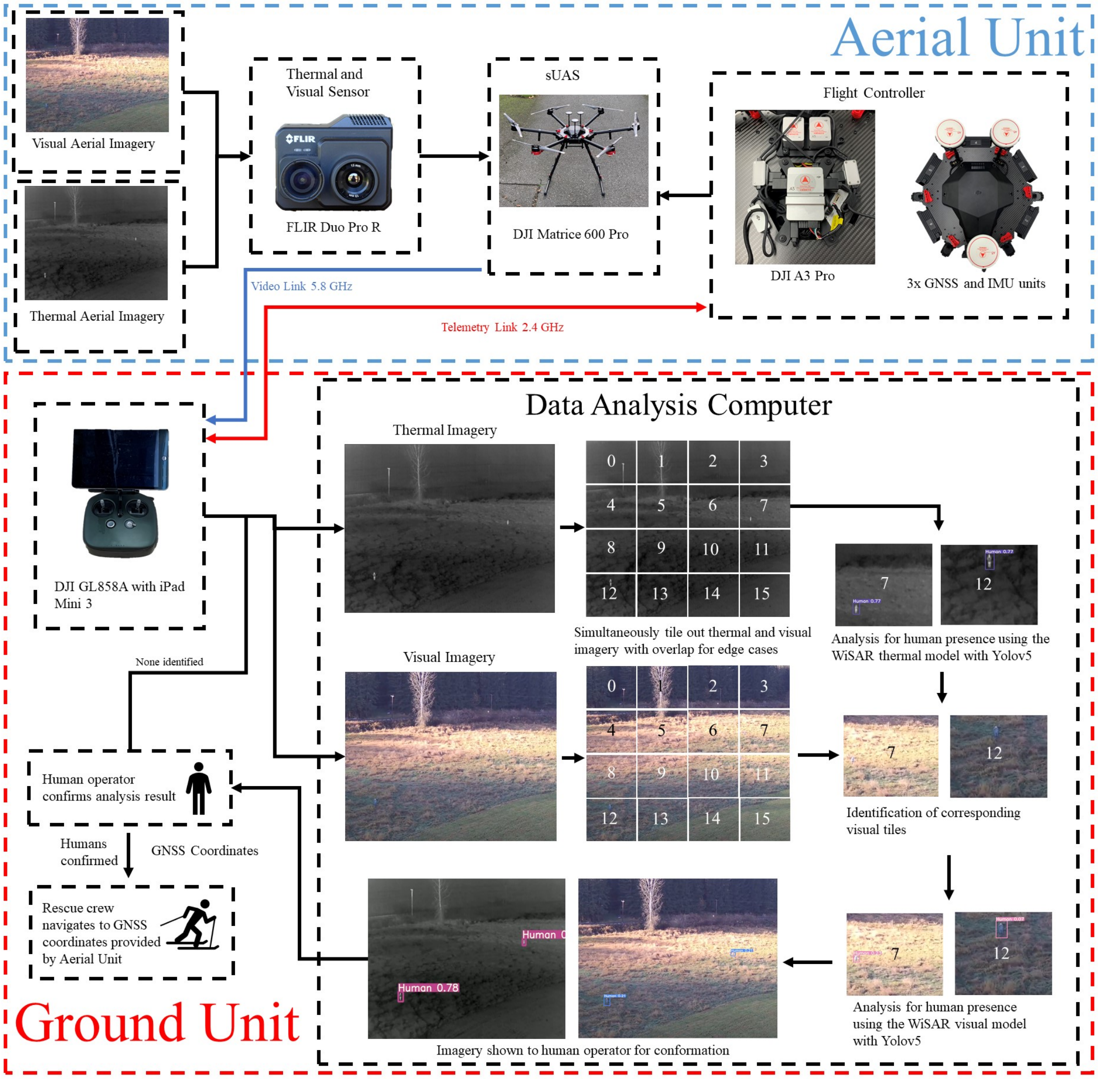

Aerial Unit
Visual Aerial Imagery
Thermal Aerial Imagery
Thermal and Visual Sensor
FLIR Duo Pro R
sUAS
DJI Matrice 600 Pro
Flight Controller
DJI A3 Pro
3x GNSS and IMU units
Video Link 5.8 GHz
Telemetry Link 2.4 GHz
Data Analysis Computer
DJI GL858A with iPad Mini 3
Thermal Imagery
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Simultaneously tile out thermal and visual imagery with overlap for edge cases
7
12
Analysis for human presence using the WiSAR thermal model with Yolov5
None identified
Visual Imagery
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Identification of corresponding visual tiles
Human operator confirms analysis result
Humans confirmed
GNSS Coordinates
Rescue crew navigates to GNSS coordinates provided by Aerial Unit
Human 0.78
Imagery shown to human operator for conformation
Analysis for human presence using the WiSAR visual model with Yolov5
Ground Unit

Fig. 4 System Architecture.

B. Software Architecture

A major challenge with object detection is balancing between computational efficiency and an acceptable accuracy threshold. Building upon the work of Benezeth et al., we will introduce two potential approaches to human detection incorporating radiometric data. The first is a two-pass system where data with low radiometric activity is culled before being fed into a conventional visual detector, decreasing computational workload. The second is a model trained on a statistical transform of the radiometric data itself, which offers substantial improvements in accuracy compared to training directly on the radiometric data.

1. Radiometric Culling

Running YOLO on a full-resolution image is impractical (further discussed in [33]). This challenge is what motivated the concept of radiometric culling.

We begin with the same basic algorithm presented in Section II of [35], assuming that radiometric background can be accurately modeled with a single Gaussian distribution, defined as [35]:

$$\mu_t(x, y) = (1 - \alpha)\sigma_{t-1}(x, y) + \alpha I_t(x, y) \tag{1}$$

$$\sigma_t^2(x, y) = (1 - \alpha)\sigma_{t-1}(x, y) + \alpha(I_t(x, y) - \mu_t(x, y))^2 \tag{2}$$

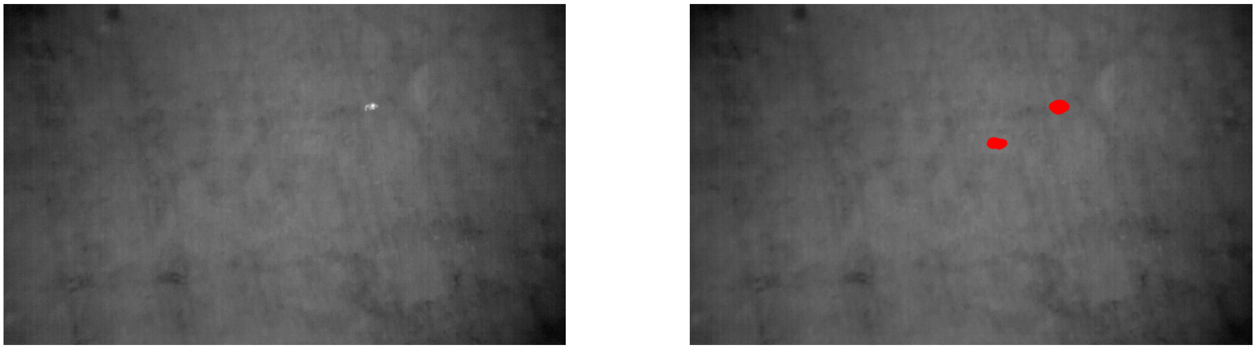
Where $I_t(x, y)$ represents the radiometric intensity of a pixel at time t, $\mu_t(x, y)$ at (x, y) pixel coordinates, and $\sigma_t(x, y)$ are the per-pixel mean and standard deviation of the Gaussian background respectively, and α is a linear interpolation parameter which we set to 0.001 from empirical observation. For $t = 0$, we set σ and μ to the actual mean and standard deviation of the radiometric data for all pixels. After that, the mean and standard deviation of each pixel are updated while taking into account previous frames using Equations 1 and 2. At any given time, we compute the per-pixel standard score:

$$z_t(x, y) = \frac{I_t(x, y) - \mu_t(x, y)}{\sigma_t(x, y)} \tag{3}$$

And then, as proposed by [35], binarize each pixel as follows:

$$B_t(x, y) = \begin{cases} 1 & \text{if } z_t(x, y) > \tau \wedge I_t(x, y) > \beta \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Where β represents a minimum temperature threshold. τ is the number of standard deviations above the Gaussian background a pixel must be to be classified as salient. Again, based on empirical observation we set $\tau = 2.5$ (the same as what [35] recommends). While this technique is effective, running it on field data yields noisy binarized images. To combat this, we run a morphological transform on the image, applying an opening operation using a 4 × 4 elliptical kernel, removing small bright spots [36]. The result is a set of coordinates corresponding to peaks of radiometric activity (Fig. 5)



(a) Input FLIR image in ideal conditions with one person.

(b) Peaks found relative to Gaussian background (shown in red on top of the original image shown in (a)).

Fig. 5 Radiometric peak detections.

As in pre-training, we subdivide the image into $n \times n$ rectangular tiles—we then apply a linear transform on the coordinates to convert them to the same space as the visual data, then cull any tile without a peak. The remaining tiles are then fed into the YOLO object detection system running our trained model. Since inference is computationally intensive, and the initial low-pass filter is cheap in comparison, this drastically decreases the computational load needed to run a model without downscaling. In ideal conditions with low radiometric background, this reduces the total number of pixels that inference is run on to a fraction of what it otherwise would be (Fig. 6).

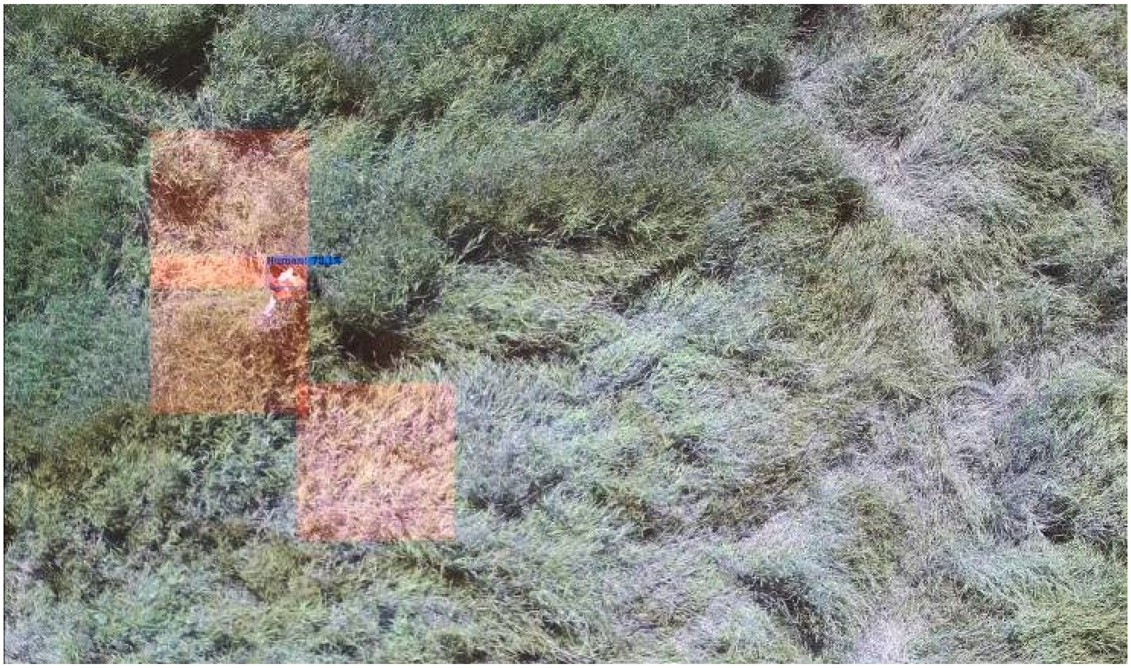


Fig. 6 Radiometric culling (active tiles shown in orange).

2. Gaussian-Background Transform

While YOLO and similar object detection frameworks operate on a per-frame basis, considering each image in isolation, it is oftentimes desirable to also consider frames relative to each other. Instead of simply training on a model on raw radiometric data, here we propose both training and inference on a transform of the data using the same Gaussian model as above. Because this transformation indirectly encodes not only the brightness of pixels, but the brightness of pixels relative to the aggregate brightness of past pixels, this effectively makes our model context-sensitive.

First, we compute a scaling factor $S_t(x, y)$ by mapping $z_t(x, y)$, defined the same as above, to the range $[0, 1]$ via a sigmoid:

$$S_t(x, y) = \frac{1}{1 + e^{-k(z_t(x,y)-z_0)}} \tag{5}$$

Where k and z_0 are constants which represent the logistic curve's steepness and midpoint, respectively. Based on empirical testing, we let $k = 2.5$ and $z_0 = 2$ for good results.

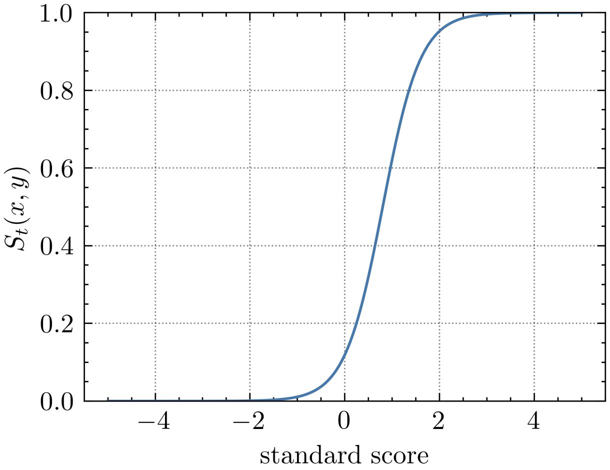


Fig. 7 The logistic scaling function.

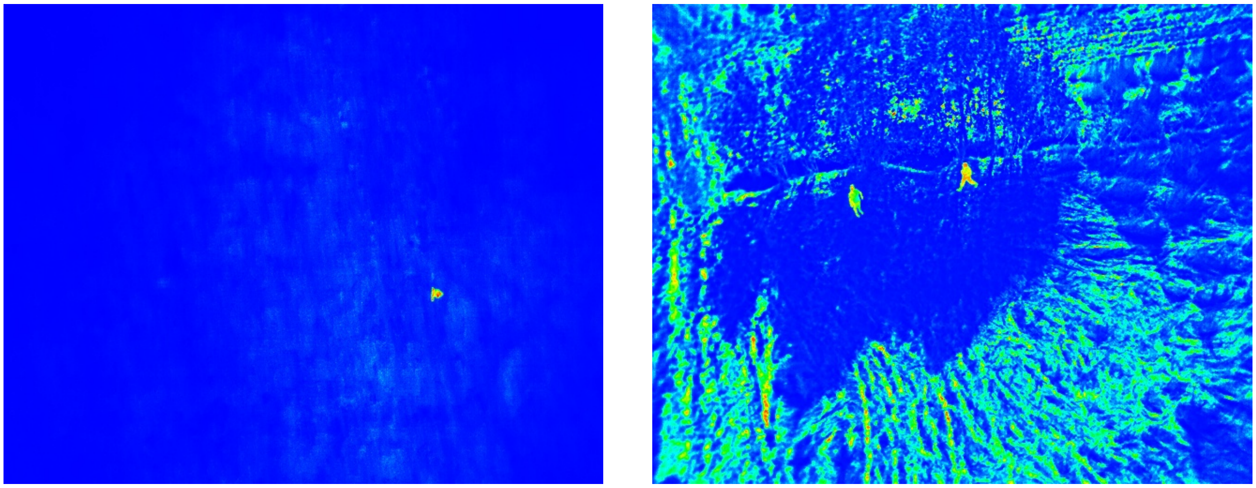


Fig. 8 Gaussian-Background HSV transformed images (transformed3). Ideal conditions can be seen on the left, and less ideal conditions on the right.

For each pixel in the radiometric image, we now multiply its raw radiometric value by the value of the scaling factor at that point—informally, we are decreasing the intensities of pixels which have not become substantially brighter, assuming that a person is substantially warmer than their surrounding environment. To increase contrast within the image, we apply a min–max normalization:

$$I_{tn}(x, y) = \frac{I_t(x, y) - I_{\min}}{I_{\max} - I_{\min}} \tag{6}$$

Finally, we translate the normalized intensity values into an HSV image based on the following criterion:

$$\begin{cases} H(x, y) = (1 - I_{tn}(x, y))\, 240^\circ \\ S(x, y) = 1 \\ V(x, y) = 1 \end{cases} \tag{7}$$

Where $H(x, y)$, $S(x, y)$, and $V(x, y)$ represent hue, saturation, and value respectively. We map the intensity along the hue from 240° (blue) to 0° (red), while fixing saturation and value at 100%. We avoid using all 360° of hue, because that would make very high values and very low values similar in color. One advantage of this HSV mapping over greyscale is that it offers 4 times more granularity, with 1020 possible values over 240° as opposed to 255 possible values.

The final output of this transform can be seen in Fig. 8. In the results presented in Section V, this transform is referred to as "transformed3".

3. Software Workflow

Training was done on AFSL Rig 1 shown in Fig. 18 whose specifications are listed in Table 4, running Ubuntu 20.04. To train and run our model, we are using the YOLOv5 framework [5], which uses the PyTorch library. We have created several Python scripts utilizing the OpenCV library to tile and cull the data for training as mentioned in Section III.B. We have scripts in place to integrate YOLOv5 with various Python applications, including a GUI intended to be used on a data analysis computer during the inference phase as shown in Fig. 4.

V. Computational Results

To evaluate how well our trained thermal and visual models perform at detecting humans under different wilderness conditions and the corresponding computation speed, we first compare the testing results of four models, Visual-YOLOv5x, Visual-YOLOv5l, Visual-YOLOv5m, and Visual-YOLOv5s. These models vary by network size as seen in Table 3. The models are tested on datasets that contain images recorded with various backgrounds, camera positions, altitudes, times of day, and temperatures, as explained before. All the comparisons are conducted with four measures: precision, recall, generalized intersection over union loss, and mean average precision, described as below. We then discuss the running speed.

When evaluating the performance of the models, we are interested in the accuracy of the number, positions, widths and heights of the bounding box outputs, with the last three specifying the precise location of the human detected. There are many metrics that take these into account, and in this work we choose the following metrics that are commonly used in object detection:

- Intersection over union (IoU), defined as

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}} \tag{8}$$

  measures the overlap between the box output and the ground truth. Generalized intersection over union (Generalized IoU) loss [37] additionally considers the distances between non-intersecting boxes, which is useful to differentiate between cases where there is no overlap at all.
- The precision, defined as

$$\text{Precision} = \frac{\text{Number of True Positives}}{\text{Number of True Positives + Number of False Positives}} \tag{9}$$

  measures how accurate the outputs are considering all testing cases [38]. An IoU threshold is typically specified for this measure—0.5 is often used.
- Similarly, recall, defined as

$$\text{Recall} = \frac{\text{Number of True Positives}}{\text{Number of True Positives + Number of False Negatives}} \tag{10}$$

  shows how likely the true objects in the image frame are to be detected, with a certain IoU threshold specified.
- The average precision (AP) is the precision value at a specified recall value on the precision recall curve, and mean average precision (mAP) is the average of multiple AP at different recall values.

Table 2 Comparison of Visual and Thermal Models

Model	Precision	Recall	GIoU	mAP$_{0.5}$	mAP$_{0.5:0.95}$	Train Time (Hours)	Speed (FPS)
Visual-YOLOv5x	**0.9012**	**0.8976**	**0.02889**	**0.9277**	**0.6940**	15.25	149.25
Visual-YOLOv5l	0.8955	0.8968	0.02917	0.9260	0.6824	8.65	250.00
Visual-YOLOv5m	0.8885	0.8985	0.02950	0.9246	0.6669	5.35	384.61
Visual-YOLOv5s	0.8614	0.8954	0.03034	0.9197	0.6354	**2.67**	**588.24**
Thermal-YOLOv5x	0.9354	0.9848	0.02416	0.9783	0.6042	4.55	185.18
Raw-Scaled	0.7295	0.9264	0.03131	**0.9246**	0.4976	**2.52**	**181.81**
Raw-HSV-Scaled	0.7346	**0.9333**	0.03097	0.9237	0.4857	**2.52**	**181.81**
Transformed3	**0.7364**	0.9241	**0.02798**	0.9232	**0.5114**	**2.52**	**181.81**

The FPS in Table 2 are created from testing the trained model on the corresponding test set of imagery using a network size of 512 × 512 on AFSL Rig 1 shown in Fig. 18 and whose specifications are listed in Table 4. In Table 2, the best results among visual models and thermal models trained on processed thermal dataset are highlighted. The first four rows show that, among the visual models that are all trained on the tiled visual dataset, Visual-YOLOv5x performs the best in terms of accuracy. The fifth row presents the performance of Thermal-YOLOv5x that is trained on the raw thermal dataset. As specified in Table 1, the size of the thermal dataset is much smaller than the visual dataset. Thus, though the performance of Thermal-YOLOv5x performs much better, we did not compare visual models with Thermal-YOLOv5x. The last two rows show the thermal models that are trained on the processed thermal datasets with a smaller dataset size. For easier comparison, we also present the Raw-Scaled model that is trained on the corresponding raw thermal dataset. Based on mAP$_{0.5:0.05:0.95}$, the more strict measure, Transformed3 has promising results.

As for the training and testing speed, for visual models, YOLOv5s has the shortest training time and highest Frame Per Second. For thermal models, the speed is the same among the last three models. In Section VII.D, we provide more detailed figures entailing the testing results and comparison between models.

Table 3 Network Size for Trainined Models

Model	Layers
Visual-YOLOv5x	484
Visual-YOLOv5l	400
Visual-YOLOv5m	316
Visual-YOLOv5s	232
Thermal-YOLOv5x	484
Raw-Scaled	484
Raw-HSV-Scaled	484
Transformed3	484

A. Advantage of Radiometric Culling

Running YOLO at reasonable speeds on high-resolution visual input typically requires substantial downscaling. Calling YOLOv5's internal Python API directly, a 3840 × 2160 input video averaged around 5.2 FPS on AFSL Rig 1 as seen in Table 4. By contrast, running radiometric culling and then YOLO on only the remaining radiometrically active tiles averaged 18 FPS—a 346% improvement in speed. However, this added performance comes with the added cost of potentially missing human subjects in environments where thermal detections suffer, such as in Fig. 10a.

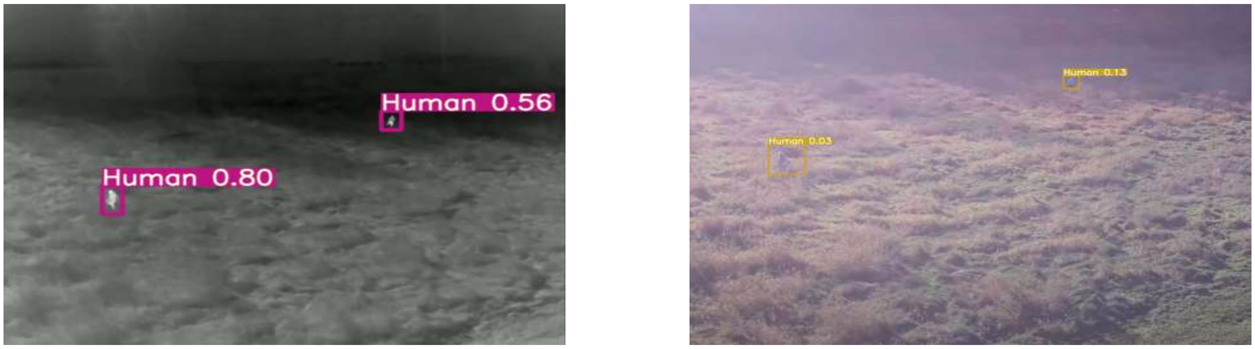
VI. Operational Results

We chose to suppress thermal and visual detections below a confidence threshold of 40%. While smaller thresholds let in false positives and larger thresholds create propensity for false negatives, for WiSAR applications the latter is disastrous, so the threshold was chosen with a bias towards allowing in false positives, which matter relatively little. Since there is a large variety of environments that can be encountered during WiSAR operations, the usefulness of visual data and radiometric data can vary significantly—the latter excels in night-time environments where there is relatively little ambient heat sources, while the former depends on high-visibility environments. Throughout this section, we will use the Thermal-YOLOv5x model to analyze thermal imagery and the Visual-YOLOv5x model to analyze visual imagery (see Table 2).

A. Daytime

During daytime operations, we found that major challenges in our image analysis occurred from variances in ambient temperature, lens glare from either the sun or reflective ground surfaces, and transitions between shady and sunny environments.

1. Lens Glare



(a) Thermal Lens Glare: From left to right the detections are 80% and 56% confident

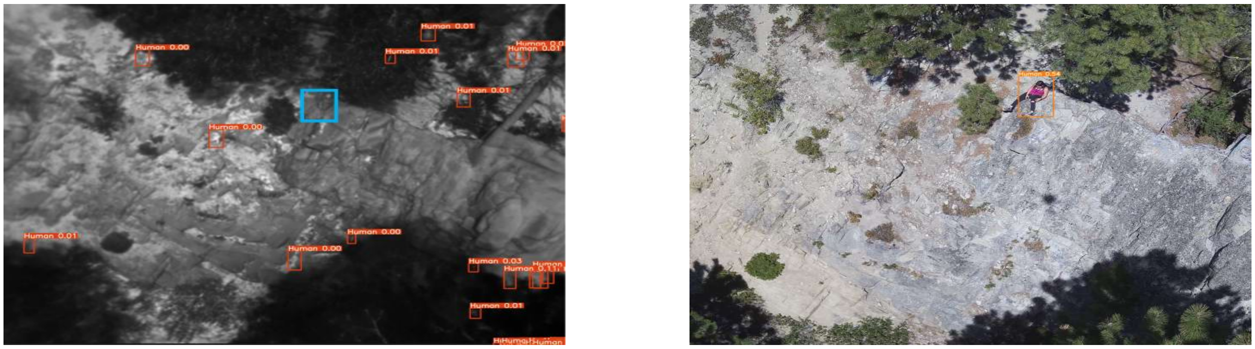
(b) Visual Lens Glare: From left to right the detections are 3% and 13% confident

Fig. 9 Examples of Lens Glare in Thermal and Visual Imagery.

The position of the sun caused lens glare to occur, seen in Fig. 9b. This is realized as a hazy appearance to the image, resulting in reduced performance in the visual detections but the thermal detections remain largely unaffected. To create Fig. 9a we used a threshold of 40% and in Fig. 9b we used a threshold of 1% for visualization purposes. The detections seen in Fig. 9b are significantly below the nominal 40% threshold used for detections and are unusable as too many false positives would be let through if used on additional images. We found that these situations occurred quite often during dusk and dawn situations when the sun was closer to the horizon. In most of these cases, the thermal detections are unaffected due to the different material used in the construction of the thermal lens, seen in Fig. 9a.

2. Extreme Temperatures

Temperature changes in the environment largely only affected thermal detections.

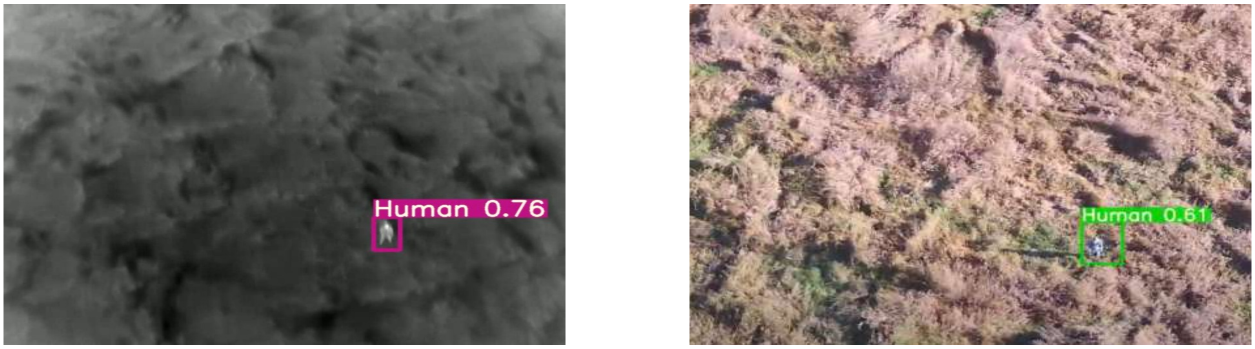


(a) Hot Thermal example where detections in this image are all false positives at or below 1% confident. The blue box in this image is for the benefit of the reader to locate where the human is. It is not a detection from our algorithm.

(b) Hot Visual with detection at 54% confidence

Fig. 10 Examples of Hot temperature (38 °C)**.**

In Fig. 10a there are no true positives present using a 1% threshold. Meaning at 38 °C, thermal offers very little to no benefit. This occurs since the background of the image is already near the body temperature of a human subject resulting in very little noticeable features. In Fig. 10b the detection at 54% is above the 40% threshold and remains useful.



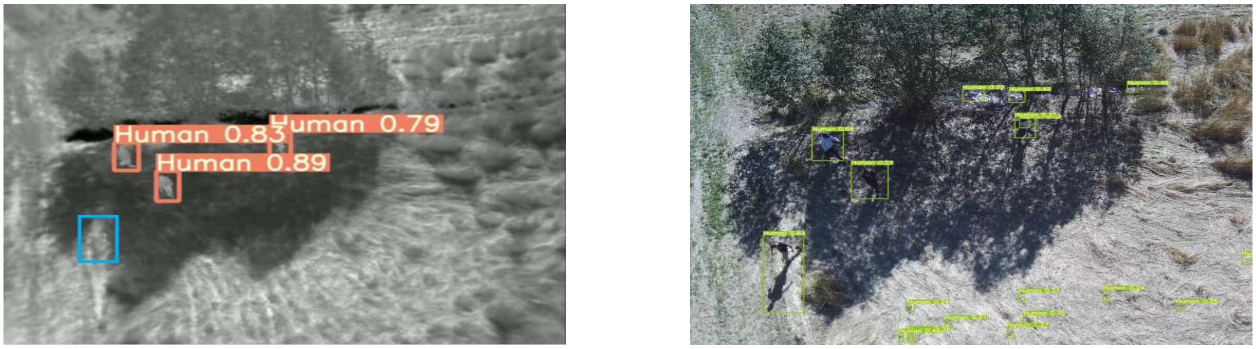
(a) Cold Thermal with detection at 76% confident

(b) Cold Visual with detection at 61% confident

Fig. 11 Examples of Cold Temperature (9 °C)

Given the much cooler background temperature, the thermal imagery has significantly more noticeable features around the human which allow for much higher detections as seen in Fig. 11a. The visual remains largely unaffected by the temperature change in Fig. 11b compared to Fig. 10b.

3. Transitions between Sunny and Shady Environments



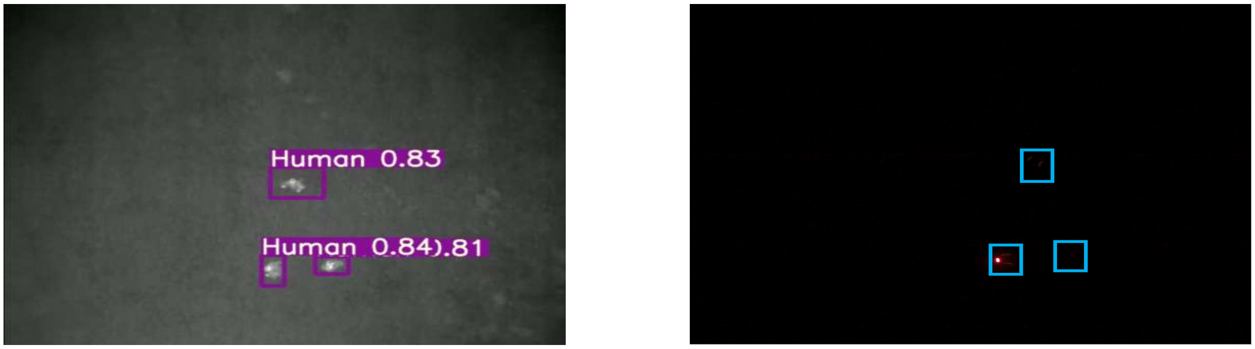
(a) Shade Sun Transition Thermal with detections from right to left at 83%, 89%, and 79% confident. The blue box in this image is for the benefit of the reader to locate where the human is. It is not a detection from our algorithm.

(b) Shade Sun Transition Visual with detections on the humans from right to left at 63%, 64%, 23%, and 37% confident.

Fig. 12 Shade Sun Transition

Transitions between sunny and shady environments proved to be a challenge because the thermal detections performed very well in shady environments but very poorly in direct sunlight at 32 °C. This is because thermal sensor is actually picking up a lack of heat in a specific area due to that specific area being kept from direct sunlight which visually appears to be a shadow in Fig. 12a. The humans in this environment had very well defined features because they were significantly hotter then the ground in the shadow. Since this environment was 32 °C, humans in the area not in the shadow of the tree had very poorly defined features resulting in a very low confidence level. The threshold was set to 20% in Fig. 12a for visualization purposes. At this threshold false positives begin to appear in the surrounding environment. The visual detection performance was less then the thermal detection performance in the shade but better in direct sunlight seen in Fig. 12a and Fig. 12b.

B. Night



(a) Night Thermal with detections from right to left at 84%, 83%, and 81% confident.

(b) Night Visual: The blue boxes in this image is for the benefit of the reader to locate where the human is. It is not a detection from our algorithm.

Fig. 13 Examples of Night

Thermal detections out perform visual detections quite dramatically in night environments seen in Fig. 13a and Fig. 13b. A threshold of 5% was used in the analysis of Fig. 13b and 40% in Fig. 13a. The lack of any visual light makes visual detections unusable. However since the environment cools off during then night compared to the day thermal detections are unaffected by the lack of visual light as seen in Fig. 13a compared Fig. 11a.

VII. Future Work

The systems in this paper form a broad foundation which is well-suited for future extension. In particular, the CV algorithms presented are versatile enough to be expanded for use with a variety of more sophisticated hardware systems.

A. Multi-Agent Systems

While range extenders offer a simple and obvious solution to handling range limitations, a more potentially interesting solution that the AFSL is currently working on is the utilization of a secondary sUAS as a communications relay point. This would not only allow for the primary sUAS to travel further away from the ground station, but also allow for flight into beyond-visual-line-of-sight conditions.

A secondary sUAS unit could also be incorporated as an effective low pass filter. By flying at a high altitude, it could classify regions of interest which a lower altitude sUAS unit could then inspect in further detail. A multi-altitude system such as this could also be implemented with a manned helicopter instead.

B. Non-multirotor sUAS

In the case of a multi-agent system, a high-altitude low-pass sUAS could effectively operate without the ability to hover in place–as a result, it could be replaced with a higher efficiency fixed-wing system. In addition, electronic vertical takeoff and landing (eVTOL) systems could be used to exploit the benefit of better efficiency and speed while maintaining the ability to hover in place.

C. On-board Computation

Large advancements in edge computing have opened up the possibility for on-flight computation. The AFSL is working on integrating a Nvidia AGX Xavier onto our current flight platform as seen in Fig. 16. This will alleviate the need to maintain a high bit-rate video link to the ground station, allowing for the expansion of the range of operation of a sUAS by a substantial amount.

D. Autonomous Path Planning

The AFSL is also working on implementing intelligent path planning algorithms, such as [10, 11]. Combined with on-flight computation, these would allow for the creation of fully autonomous WiSAR platforms, limited only by on-board battery life.

Appendix

Appendix A: WiSAR Bill

Detailed Expense Report for SAR 11c05025 / State SAR #11-1616 (Thompson)

Completed by Sgt Kent Sisson

ST Hours = Straight-time Hrs / OT = Overtime Hrs
ST Rate = Straight time rate (not to include benefits)
OT Rate+ = Overtime rate + benefits (insurance, FICA, LI)
VOL = CCSO Volunteer

Day 1 Personnel Cost				1-Jun-11			
Personnel	Position	ST Hours	OT Hours		ST Rate	OT Rate+	Total
Sisson	ATV Ops	8	2		$38.92	$81.99	$475.34
Ellis	ATV Ops	10	0		$31.73	$70.19	$317.30
Bryant	Support	8	2		$20.58	$41.70	$248.04
							$1,040.68
Day 2 Personnel Cost				2-Jun-11			
Personnel	Position	ST Hours	OT Hours		ST Rate	OT Rate+	Total
Sisson	I/C	8	7.5		$38.92	$81.99	$926.28
Ellis	Ops Chief	10	5.5		$31.73	$70.19	$703.34
Bryant	Log. Chief	8	7.5		$20.58	$41.70	$477.39
Seabright	Team 1 Lead	0	15		N/A	$59.79	$896.85
Nesary	Team 1	0	15		N/A	$64.59	$968.85
Norton	Team 1	0	15		N/A	$62.57	$938.55
Huddleston	Team 2 Lead	10	6		$35.30	$77.63	$818.78
Moran	Team 2	Vol	Vol		$0	$0	$0
Schively	Team 2	Vol	Vol		$0	$0	$0
Shales	Pilot Air 20	9	6		$25.00	$50.00	$525.00
Lawrence	TFO Air 20	12	2		$30.22	$68.88	$500.40
Wisemore	Admin/PIO	8	2		$37.70	$46.13	$383.86
Agnew	PIO Lead	10	2		$42.44	$80.46	585.32
Isaacson	Logistics	Vol	Vol		$0	$0	$0
Coffman	Logistics	Vol	Vol		$0	$0	$0
							$7,724.62
Additional Expenses:							
	Day 1	Day 2					Total
Meals	$0	$54.00					$54.00
Aircraft Fuel	$0	$398.60	(100 gal)				$398.60
Aircraft Ops	$0	15hrs @	$100hr				$1,500
							$1,952.60
			Total Expense of SAR =				$10,717.90

Appendix B: Hardware

Aerial Unit

sUAS

The particular sUAS model selected was a DJI Matrice 600 Pro. It has a 9.5 kg weight without the WiSAR payload. It is 114 cm motor to motor with the arms unfolded with 54 cm props on each of the six motors and stands 55 cm tall. It uses six DJI TB47s mAh 6S batteries. A DJI A3 Pro is used to control the sUAS. It is connected to a DJI Lightbridge 2 which uses 2.4 GHz for telemetry and 5.8 GHz for video transmission at 1920 × 1080 at 60 FPS with an unobstructed range of 5 km. There are three GNSS/IMU modules for redundancy (standard for DJI A3 Pro) which allows for ±1.5 m horizontally and ±0.5 m vertically of accuracy. This aircraft was chosen for its ability to hold up to 6 kg max payload which allows for a lot of freedom in the payloads which could be experimented with. It also requires minimal effort in integrating, as the transmitter, a DJI GL858A, includes HDMI and SDI video out ports, which allows for the video to be piped into the data analysis computer.

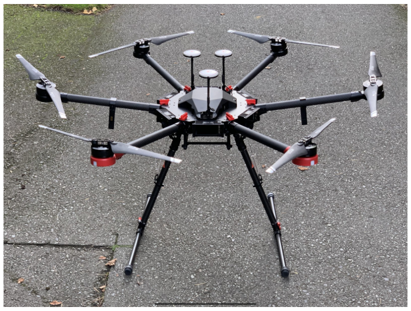
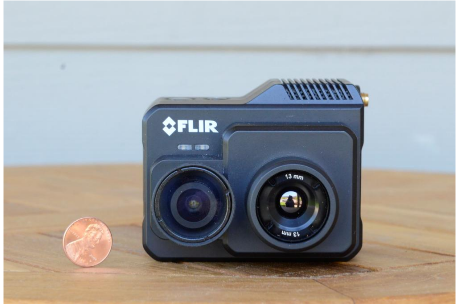


Fig. 14 DJI Matrice 600 Pro



)

Fig. 15 FLIR DUO PRO R

Payload

The main sensor used is a FLIR Duo Pro R with a weight of 325 g. The spectral band of the uncooled VOx microbolometer sensor is 7.5 µm to 13.5 µm with a resolution of 640 × 512 and a FOV of 45° × 37°. The visual sensor has a resolution of 4000 × 3000 and a FOV of 56° × 45°. The HDMI output provides a 1920 × 1080 video feed to the Lightbridge unit. A secondary FlySky FS-i10 transmitter and Turnigy TGY-iA10b receiver is used to interact with the FLIR via the three pulse-width modulation (PWM) connections to the 10-pin accessory port. This allows the operator to change the color space used for the thermal image and select which video feed (thermal and visual) is being sent out via HDMI. The FLIR is mounted on a Gremsey T3V3 Gimbal seen in 16. Power for the FLIR is provided from the 14.5 V output at the bottom of the gimbal which interfaces with the accessory port. The FLIR Duo Pro R was chosen for its inclusion of both a thermal and visual sensors as they are both integral in this research.

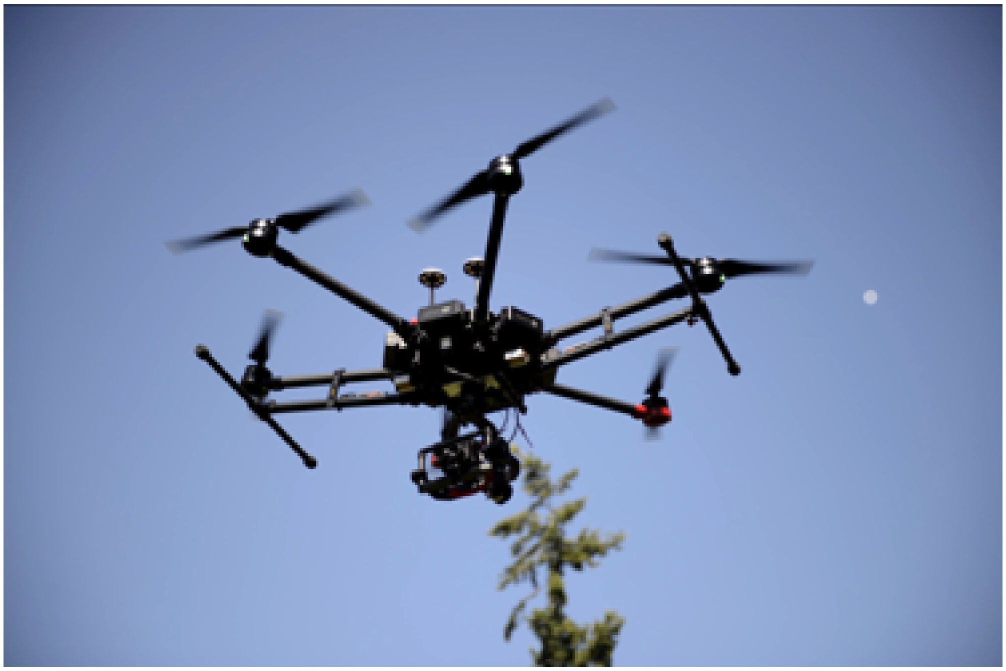


Fig. 16 Full Platform

Ground Unit

Video Capture Card

The external video capture card used is Razer Ripsaw HD which is able to support up to 1920 × 1080 at 60FPS. An internal video card could be used as well but an external one was ultimately selected as we wanted the additional modular functionality. This particular model was chosen for its ability to support the max resolution and frame rate that the DJI Lightbridge 2 module supports.

Fig. 17 Razer Ripsaw HD

Fig. 18 AFSL Rig 1 Data Analysis Computer

Data Analysis Computer

The data analysis computer provides the necessary computing power required to analyze the incoming data from the capture card in real time. The main visual image analysis is done by the YOLOv5 neural network in the PyTorch framework [5] based off of [34] and [39]. As YOLOv5, like most other real time CNNs, is primarily bottle-necked by the GPU, this computer needed to have a high-end GPU, but also remain portable and rugged enough to take to real WiSAR operations.

Table 4 Specifications of the computers tested for data analysis.

Data Analysis Computer	GPU	CPU	RAM
AFSL Rig 1	Nvidia Titan RTX 24 GB VRAM	AMD Ryzen 3900X	64 GB DDR4 3000 MHz

The GPU that we choose is the Nvidia Titan RTX, mainly for its 4608 CUDA cores and 24 GB of VRAM as this allows us to use a network image size of 512×512 at an average of 149.25 FPS on the Visual-YOLOv5x as seen in Table 2. The CPU is less important for running the CNN, but is valuable for running the numerical thermal analysis algorithms. Thus a mid range CPU such as the AMD Ryzen 3900X sufficed. The amount of RAM was also not as important so 64 GB of DDR4 at 3000 MHz was also sufficient. The chassis was another important factor of this computer as it needed to be small enough to remain portable but also rugged enough to withstand being in a car on rough terrain for longer periods of time. While there are more rugged cases, such as the Silverstone MM01 or the line of rackmount cases from Pelican, we chose to use a Corsair 110Q as we intended for this computer to be used primarily in the lab. For regular use on site, a more rugged case is recommended. Additionally, there are no HDDs used as these are less resistant to vibrations due to physical moving parts within them. NVME based M.2 SSDs are used in there place due to there small form factor, fast transfer speeds and resistance to vibrations.

Appendix C: Challenges Encountered

Video Transmission Range

One of the key pieces in our system is the HD video transmission system. We found that the distance from the operator to the drone was a significant limitation in the quality of the imagery that was received from the DJI LB2 module. We encountered frame drops at distances beyond 500 m. Even in a rural farm environment with clear line of sight this was still an issue as seen in Fig. 19.

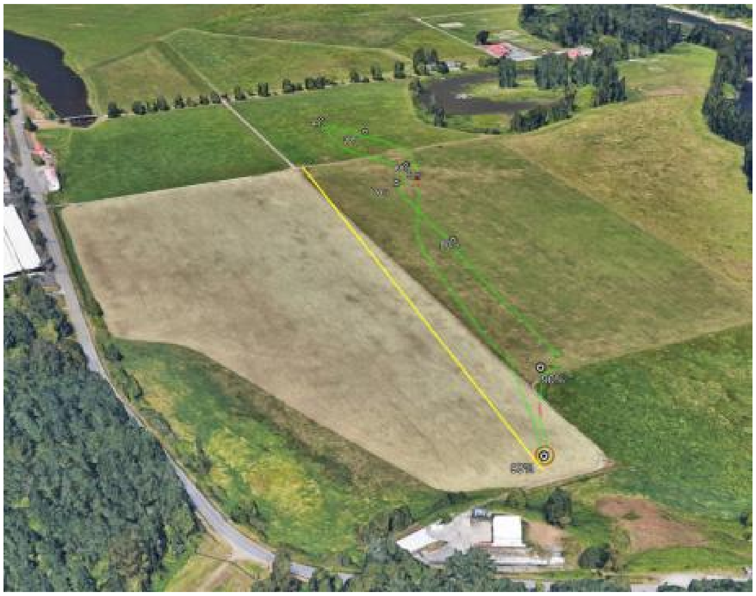
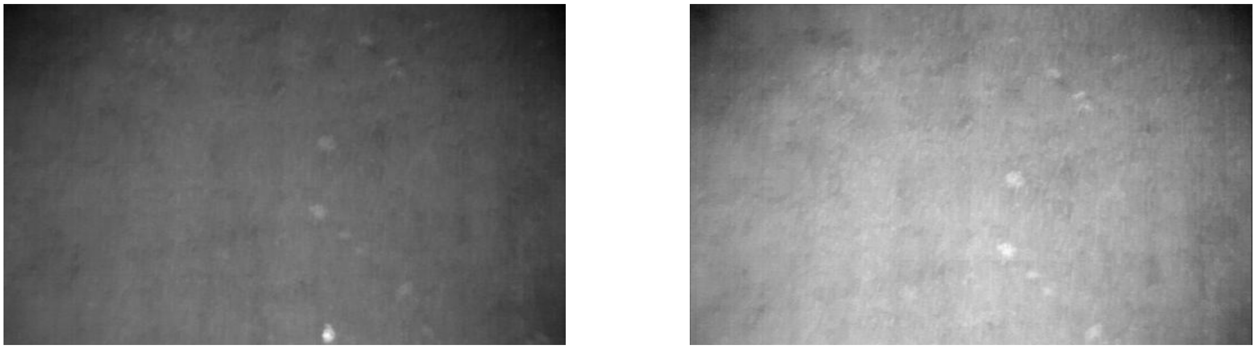


Fig. 19 The green line is the flight path of the sUAS and the yellow line is 500 m

Streaming Radiometric Data

The thermal camera we used automatically will assign a non-fixed color scale to the imagery, assigning a certain color to the coolest object in frame and another color to the hottest. If a human was in frame and the hottest color value was assigned to them but then quickly walked out of frame, the the contrast of the image would dramatically jump up because the new coolest and hottest points of the image are at a similar temperature as seen in Fig. 20a and Fig. 20b.



(a) Pre-transition

(b) Post Transition

Fig. 20 Contrast Issue in Fast Transitions

Appendix D: Additional Computational Results

In this part, we discuss mainly on the results of the best visual and thermal models, and then present the more detailed comparison for the four visual models and three thermal models.

As shown in Fig. 21, after 15 hours of training for the visual model and 4.5 hours for the thermal model, the precision grows to 90.12% and 93.54% respectively, and the recall grows to 89.76% and 98.48% respectively, when tested on testing dateset. Ideally, for the WiSAR purpose, we first prioritize reducing the number of false negative cases, thus human subjects will always be detected, followed by reducing the amount of false positive cases, avoiding false information for the human operator. Based on the results of the two measures, both of the models perform well. The comparison between visual and thermal models is less useful because the sizes of the visual and thermal datasets are different.

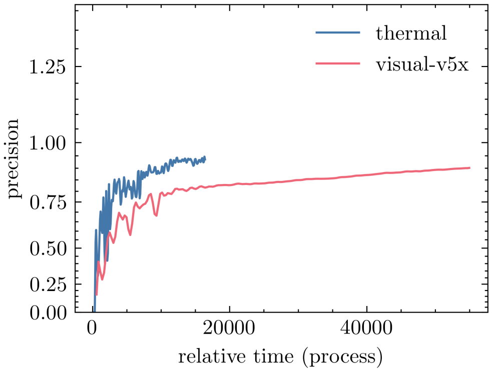


Fig. 21 Precision curve of the best thermal and visual models during training

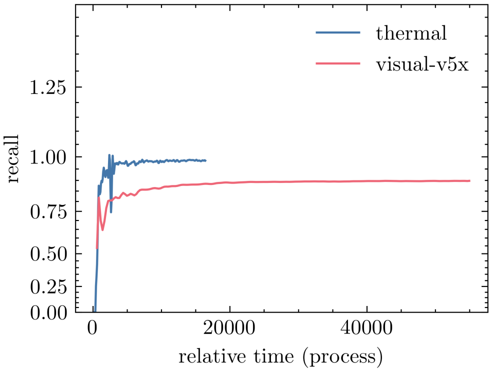


Fig. 22 Recall curve of the best thermal and visual models during training

In Fig. 23, we can see the Generalized IoU loss is reduced to 0.02889 for visual model and 0.02416 for thermal model.

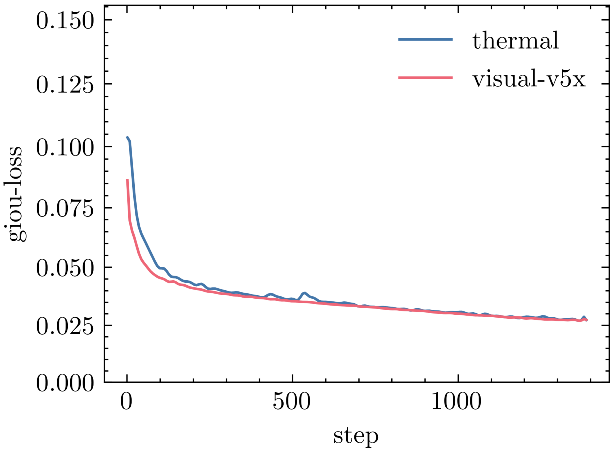


Fig. 23 Generalized Intersection over Union Loss

As shown in Fig. 24, with recall being 0.50, both models have high mAP. Figure 25, which shows a more challenging measure that considers AP with recall ranging from 0.50 to 0.95 with step 0.05, is further indicative of strong performance across the board.

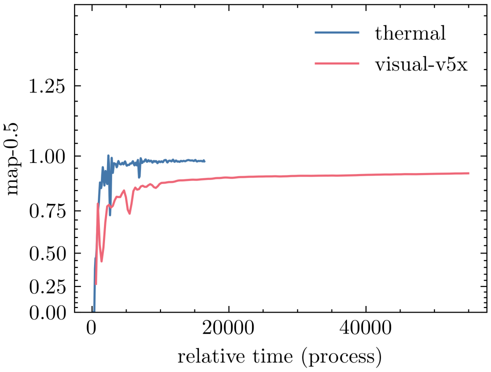


Fig. 24 mAP@0.5 curve of the best thermal and visual models during training

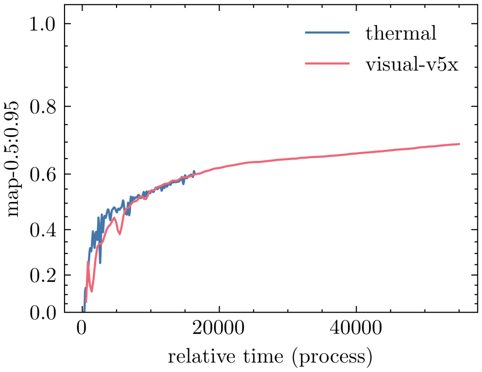


Fig. 25 mAP@0.5:0.95 curve of the best thermal and visual models during training

In Figs. 26 to 30 compare both to the performance of several other visual models. As shown, Visual-YOLOv5x is the best-performing visual model.

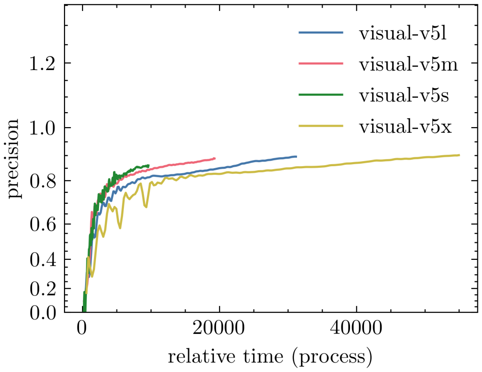


Fig. 26 Precision curve of four YOLO models during training

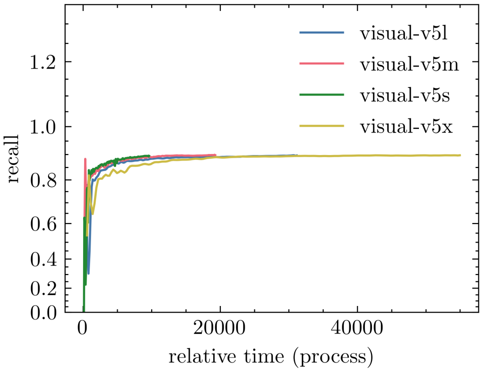


Fig. 27 Recall curve of four YOLO models during training

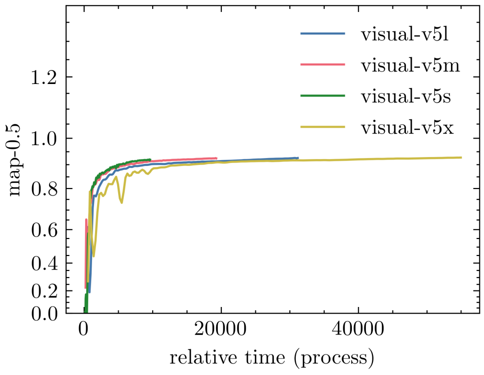


Fig. 28 MAP@0.5 curve of four YOLO models during training

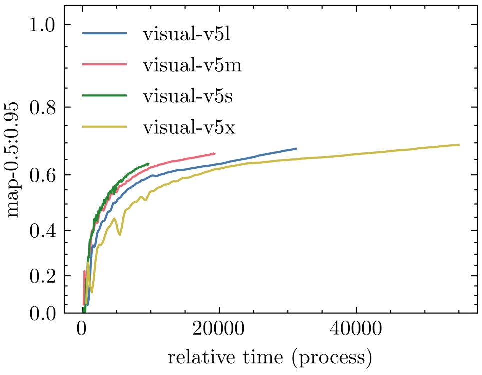


Fig. 29 MAP@0.5:0.05:0.95 curve of four YOLO models during training

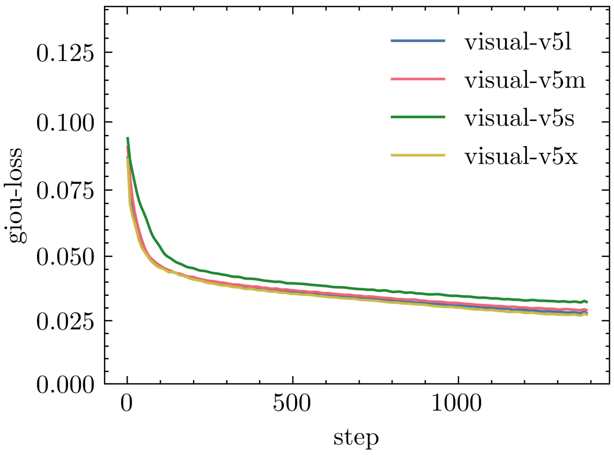


Fig. 30 GIoU curve of four YOLO models during training

Figures 31 to 35 show results of training YOLO on three forms of the thermal data from 3183 images from the UW WiSAR dataset. “thermal-raw-scaled” is an exact recreation of the video output FLIR provides. It uses min–max normalization to scale the raw temperature values to greyscale values between 0 and 255. “thermal-raw-hsv-scaled” represents this same raw temperature data, but represented by hue in HSV as described in Section IV.B.2, as opposed to in greyscale. This should theoretically provide a marginal benefit because of the wider range of available colors. “thermal-transformed3” represents the output of the complete transformation described in Section IV.B.2. Figs. 31 to 35 and Table 2 show how Transformed3 performs compared to models trained on raw greyscale images and raw HSV-converted ones. As shown, Transformed3 is the best-performing radiometric model.

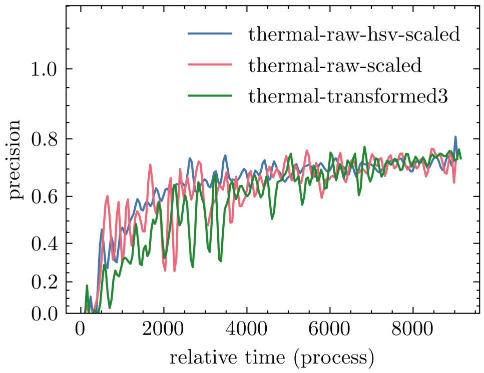


Fig. 31 Precision curve of three thermal models during training

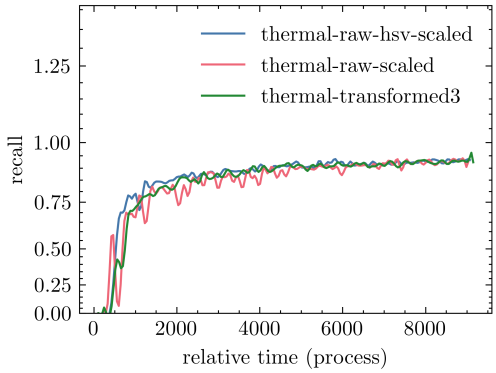


Fig. 32 Recall Curve of three thermal models during training

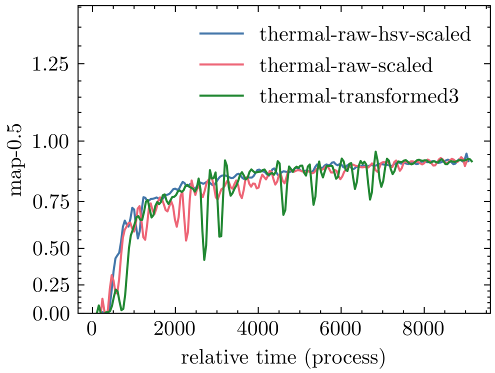


Fig. 33 mAP@0.5 curve of three thermal models during training

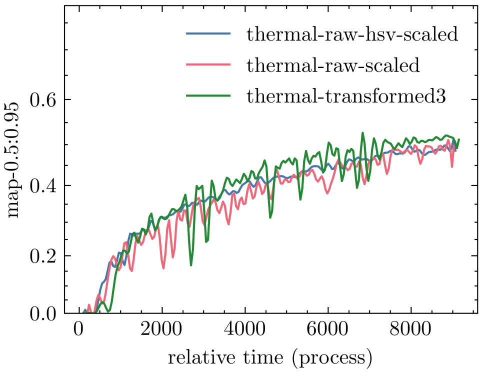


Fig. 34 mAP@0.5:0.05:0.95 curve of three thermal models during training

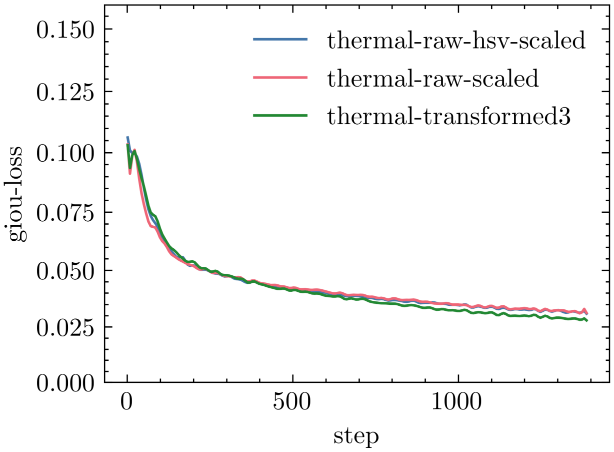


Fig. 35 GIoU curve of three thermal models during training

Acknowledgments

The authors would like to thank David Shean and the members of the Terrain Analysis and Cryosphere Observation Laboratory for providing initial computing resources for training our models and for the many great suggestions regarding the the acquisition and training of our dataset. The authors would like to thank Andy von Flotow from Hood Technology for contributions to the research vision. The Chelan County Sheriffs Office provided invaluable insight into the challenges in real WiSAR operations. Carnation Farms generously supported the flight testing of this project by allowing tests to be conducted at their facility. Jim Cronan generously provided his forest service test sites for our data acquisition flights. Many members of the UW's Autonomous Flight Systems Laboratory contributed to the research including Danny Broyles, Karina Bridgman, Helen Kuni, Saket Gollapaudi, Howard Peng, Kenneth Wilsey, Athil George, Rori Fortmann, Liam Mortell, Nicholas Orndorff, Nicholas Price, Hannah Rotta, Rostyk Svitelskyi, Hunter Warick, Eli Sitchin, Cameron Tavakoli, Conner Knoth, Ethan Uehara, Daniel Moore, Ravi Patel and An Nguyen. This research was sponsored by the Joint Center for Aerospace Technology Innovation (JCATI) as part of the project entitled "Drone aided wilderness search and locate (SAL)." The data in Section V and Appendix D are results from Weights & Biases [40].

References

[1] Rudol, P., and Doherty, P., "Human Body Detection and Geolocalization for UAV Search and Rescue Missions Using Color and Thermal Imagery," *2008 IEEE Aerospace Conference*, 2008, pp. 1–8. https://doi.org/10.1109/AERO.2008.4526559.

[2] Murphy, R. R., "Human-robot interaction in rescue robotics," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 34, No. 2, 2004, pp. 138–153.

[3] Westall, P., Ford, J. J., O'Shea, P., and Hrabar, S., "Evaluation of Maritime Vision Techniques for Aerial Search of Humans in Maritime Environments," *2008 Digital Image Computing: Techniques and Applications*, 2008, pp. 176–183.

[4] Goodrich, M. A., Morse, B. S., Gerhardt, D., Cooper, J. L., Quigley, M., Adams, J. A., and Humphrey, C., "Supporting wilderness search and rescue using a camera-equipped mini UAV," *Journal of Field Robotics*, Vol. 25, No. 1âĂŘ2, 2008, pp. 89–110. https://doi.org/10.1002/rob.20226, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.20226.

[5] Jocher, G., Stoken, A., Borovec, J., NanoCode012, ChristopherSTAN, Changyu, L., Laughing, tkianai, Hogan, A., lorenzomammana, yxNONG, AlexWang1900, Diaconu, L., Marc, wanghaoyang0106, ml5ah, Doug, Ingham, F., Frederik, Guilhen, Hatovix, Poznanski, J., Fang, J., äžŐåŁŻåĘŻ, L. Y., changyu98, Wang, M., Gupta, N., Akhtar, O., PetrDvoracek, and Rai, P., "ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements," , Oct. 2020. https://doi.org/10.5281/zenodo.4154370, URL https://doi.org/10.5281/zenodo.4154370.

[6] Lum, C., Gardner, S., Jordan, C., and Dunbabin, M., "Expanding diversity in STEM: Developing international education and research partnerships in a global society," *Proceedings of the 123rd American Society for Engineering Education (ASEE) Annual Conference and Exposition 2016: Jazzed About Engineering Education*, edited by P. A. Sanger, American Society for Engineering Education (ASEE), United States of America, 2016, pp. 7613–7629. URL https://eprints.qut.edu.au/117672/.

[7] Lum, C., Mackenzie, M., Shaw-Feather, C., Luker, E., and Dunbabin, M., "Multispectral Imaging and Elevation Mapping from an Unmanned Aerial System for Precision Agriculture Applications," *Proceedings of the 13th International Conference on Precision Agriculture*, 2016, pp. 4–10.

[8] Lum, C. W., Summers, A., Carpenter, B., Rodriguez, A., and Dunbabin, M., "Automatic Wildfire Detection and Simulation using Optical Information from Unmanned Aerial Systems," *SAE Technical Paper*, SAE International, 2015, pp. 2–3. https://doi.org/10.4271/2015-01-2474, URL https://doi.org/10.4271/2015-01-2474.

[9] Lum, C., Rysdyk, R., and Pongpunwattana, A., "Autonomous Airborne Geomagnetic Surveying and Target Identification," *Proceedings of the AIAA Infotech@Aerospace Conference*, 2005, pp. 1–12. https://doi.org/10.2514/6.2005-7039.

[10] Lum, C., and Vagners, J., "A Modular Algorithm for Exhaustive Map Searching Using Occupancy Based Maps," *Proceedings of the AIAA Infotech@Aerospace Conference*, 2009, pp. 1–22. https://doi.org/10.2514/6.2009-1839.

[11] Lum, C. W., Vagners, J., and Rysdyk, R. T., "Search Algorithm for Teams of Heterogeneous Agents with Coverage Guarantees," *Journal of Aerospace Computing, Information, and Communication*, Vol. 7, No. 1, 2010, pp. 1–31. https://doi.org/10.2514/1.44088, URL https://doi.org/10.2514/1.44088.

[12] Office of the Federal Register, "Remote pilot in command," *Code of Federal Regulations, title 14 §107.19*, 2020. URL https://gov.ecfr.io/cgi-bin/text-idx?SID=278ff16c2afa77a2460e959e39aacdbc&mc=true&node=pt14.2.107&rgn=div5#se14.2.107_119.

[13] Tso, K. S., Tharp, G. K., Zhang, W., and Tai, A. T., "A multi-agent operator interface for unmanned aerial vehicles," *Gateway to the New Millennium. 18th Digital Avionics Systems Conference. Proceedings (Cat. No.99CH37033)*, Vol. 2, 1999, pp. 6.A.4–6.A.4.

[14] Adams, J. A., Cooper, J. L., Goodrich, M. A., Humphrey, C., Quigley, M., G, B., and Morse, B. S., "Camera-Equipped Mini UAVs for Wilderness Search Support: Task Analysis and Lessons from Field Trials," , 2007.

[15] Gerhardt, D. D., "Feature-based Mini Unmanned Air Vehicle Video Euclidean Stabilization with Local Mosaics," Master's thesis, Brigham Young University, Provo, Utah 84602, 2007.

[16] Buddharaju, P., Pavlidis, I., Tsiamyrtzis, P., and Bazakos, M., "Physiology-Based Face Recognition in the Thermal Infrared Spectrum," *IEEE transactions on pattern analysis and machine intelligence*, Vol. 29, 2007, pp. 613–26. https://doi.org/10.1109/TPAMI.2007.1007.

[17] de Oliveira, D., and Wehrmeister, M., "Using Deep Learning and Low-Cost RGB and Thermal Cameras to Detect Pedestrians in Aerial Images Captured by Multirotor UAV," *Sensors*, Vol. 18, No. 7, 2018, p. 2244. https://doi.org/10.3390/s18072244, URL http://dx.doi.org/10.3390/s18072244.

[18] "Drone thermal camera - Workswell WIRIS Security," , 2020. URL https://www.drone-thermal-camera.com/products/search-and-rescue-thermal-camera-for-drone-wiris-securitu-workswell/.

[19] "FLIR Duo Pro R." , 2020. URL https://www.flir.com/products/duo-pro-r/?model=436-0345-62-0.

[20] Lienhart, R., and Maydt, J., "An extended set of Haar-like features for rapid object detection," *Proceedings. International Conference on Image Processing*, Vol. 1, 2002, pp. I–900. https://doi.org/10.1109/ICIP.2002.1038171.

[21] Mallat, S., "Understanding deep convolutional networks," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 374, No. 2065, 2016, p. 20150203. https://doi.org/10.1098/rsta.2015.0203, URL http://dx.doi.org/10.1098/rsta.2015.0203.

[22] Lawrence, S., Giles, C. L., Ah Chung Tsoi, and Back, A. D., "Face recognition: a convolutional neural-network approach," *IEEE Transactions on Neural Networks*, Vol. 8, No. 1, 1997, pp. 98–113.

[23] Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E., "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, Vol. 2, No. 1, 2015, p. 1.

[24] Abdelouahab, K., Pelcat, M., Serot, J., and Berry, F., "Accelerating CNN inference on FPGAs: A survey," *arXiv preprint arXiv:1806.01683*, 2018.

[25] Bejiga, M. B., Zeggada, A., Nouffidj, A., and Melgani, F., "A convolutional neural network approach for assisting avalanche search and rescue operations with UAV imagery," *Remote Sensing*, Vol. 9, No. 2, 2017, p. 100.

[26] Gotovac, S., PapiÄĞ, V., and MaruÅąiÄĞ, Å., "Analysis of saliency object detection algorithms for search and rescue operations," *2016 24th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2016, pp. 1–6.

[27] Portmann, J., Lynen, S., Chli, M., and Siegwart, R., "People detection and tracking from aerial thermal views," *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 1794–1800.

[28] Rasmussen, N. D., Morse, B. S., Goodrich, M. A., and Eggett, D., "Fused visible and infrared video for use in Wilderness Search and Rescue," *2009 Workshop on Applications of Computer Vision (WACV)*, 2009, pp. 1–8.

[29] Ju Han, and Bhanu, B., "Detecting moving humans using color and infrared video," *Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, MFI2003.*, 2003, pp. 228–233.

[30] Bochkovskiy, A., "Yolo_Mark," https://github.com/AlexeyAB/Yolo_mark, 2019.

[31] Zhu, P., Wen, L., Bian, X., Ling, H., and Hu, Q., "Vision meets drones: A challenge," *arXiv preprint arXiv:1804.07437*, 2018.

[32] Zhu, P., Wen, L., Du, D., Bian, X., Hu, Q., and Ling, H., "Vision Meets Drones: Past, Present and Future," *arXiv preprint arXiv:2001.06303*, 2020.

[33] Ozge Unel, F., Ozkalayci, B. O., and Cigla, C., "The Power of Tiling for Small Object Detection," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019, pp. 4–6.

[34] Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M., "YOLOv4: Optimal Speed and Accuracy of Object Detection," , 2020.

[35] Benezeth, Y., Emile, B., Laurent, H., and Rosenberger, C., "A Real Time Human Detection System Based on Far Infrared Vision," *Image and Signal Processing*, edited by A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 76–84.

[36] Delmas, P., "Morphological Image Processing," https://www.cs.auckland.ac.nz/courses/compsci773s1c/, 2006.

[37] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S., "Generalized intersection over union: A metric and a loss for bounding box regression," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 658–666.

[38] Mehdiyev, N., Enke, D., Fettke, P., and Loos, P., "Evaluating Forecasting Methods by Considering Different Accuracy Measures," *Procedia Computer Science*, Vol. 95, 2016, pp. 264–271. https://doi.org/10.1016/j.procs.2016.09.332.

[39] Redmon, J., and Farhadi, A., "YOLOv3: An Incremental Improvement," , 04 2018.

[40] Biewald, L., "Experiment Tracking with Weights and Biases," , 2020. URL https://www.wandb.com/.