Search and Rescue with Airborne Optical Sectioning

David C. Schedl*, **Indrajit Kurmi***, **and Oliver Bimber***

*first.lastname@jku.at; Johannes Kepler University, Faculty of Engineering and Natural Sciences, Linz, 4040, Austria

ABSTRACT

We show that automated person detection under occlusion conditions can be significantly improved by combining multi-perspective images before classification. Here, we employed image integration by Airborne Optical Sectioning (AOS)—a synthetic aperture imaging technique that uses camera drones to capture unstructured thermal light fields—to achieve this with a precision/recall of 96/93%. Finding lost or injured people in dense forests is not generally feasible with thermal recordings, but becomes practical with use of AOS integral images. Our findings lay the foundation for effective future search and rescue technologies that can be applied in combination with autonomous or manned aircraft. They can also be beneficial for other fields that currently suffer from inaccurate classification of partially occluded people, animals, or objects.

In 2018, 2 723 search-and-rescue (SAR) incidents were reported by the National Park Service Units in 165 national parks throughout the United States. The operation costs added up to $4.5 M, 36% of which were related to air operations. In the same year, 2 292 alpine SAR operations were performed by ÖAMTC air emergency helicopters in Austria, and 2 597 SAR missions were carried out by helicopters in the United Kingdom. In the UK 461 (18%) of these flights searched for persons or crafts.

Rescuing, lost, ill or injured persons often involves searching densely forested terrain. Sunlight is mostly blocked by trees and other vegetation, and the forest ground reflects little light. Thermal imaging systems are therefore employed to visualize the temperature difference between human bodies and the surrounding environment. Autonomous unmanned drones will increasingly replace manned helicopters in future SAR operations,[1,2] as they offer higher flexibility at lower cost. As in autonomous driving,[3,4] this requires for robust automatic people detection mechanisms.

However, such search missions remain challenging due to occlusion and high heat radiation by trees under direct sunlight. Figure 1 illustrates examples of thermal images of two different forest types (mixed and conifer forests)—captured from a drone—in which people on the ground can barely be detected because *(i)* their heat footprint is largely occluded by trees and *(ii)* the temperature of sunlight reflected by branches and tree crowns appears similar to body temperature on sensors. Obviously, simply thresholding the heat signal will not enable person detection.

Synthetic apertures (SA) approximate the signal of a single hypothetical wide aperture sensor by means of either an array of static small aperture sensors or a single moving small aperture sensor whose individual signals are computationally combined to increase resolution, depth-of-field, frame rate, contrast, and signal-to-noise ratio. This principle has been used in fields such as radar,[5–7] radio telescopes,[8,9] interferometric microscopy,[10] sonar,[11,12] ultrasound,[13,14] LiDAR,[15,16] and imaging.[17–24]

With Airborne Optical Sectioning (AOS)[25–29], we have introduced a synthetic-aperture imaging technique that uses a camera drone to capture an unstructured light field (i.e., a set of single images at unstructured sampling positions, Fig. 1; further details in Section S1 of the Supplementary Material). Color and thermal images that are recorded within the shape of a wide synthetic aperture area above a forest are combined (registered and integrated) to computationally remove occluders, such as trees and other vegetation. Applied to thermal imaging, our technique makes the radiated heat signal of largely occluded targets (e.g., a human body hidden in dense undergrowth) visible by integrating multiple thermal recordings from slightly different perspectives. The outcome is a mostly occlusion-free view of the forest ground.

Our hypothesis is that AOS integral images will enable hitherto unfeasible automated person detection in dense forests. Initial field experiments, such as that shown in Fig. 2, corroborate our hypothesis. While the heat footprint in single thermal recordings shows mostly a random pattern due to varying partial occlusion from different views, AOS results often reveal the recognizable shape of a human. As mentioned above, considering only the strength of the heat signal is insufficient for detection, as similar heat signals are produced by direct reflection of sunlight from the trees.

In this article we show that the detection rate can be significantly improved by combining multiple images (i.e., by registering and integrating) before detection rather than by combining multiple detection results from individual images. In our experiments, we achieve an average precision (AP) score of 92.2%, compared to an AP score of 24.8% with single images. Our findings pave the way for future autonomous SAR technologies that focus on finding lost and injured people in dense forests. Since fast operation is critical to such missions, computer-supported analysis of the enormous amount of image data is essential. However, human detection by means of AOS requires a specialized training dataset. Multi-spectral datasets that are available for autonomous driving,[30,31] for example, cannot be applied, as they mainly contain upright (i.e., standing, walking or running) people in urban environments and do not include AOS-specific optical aberrations

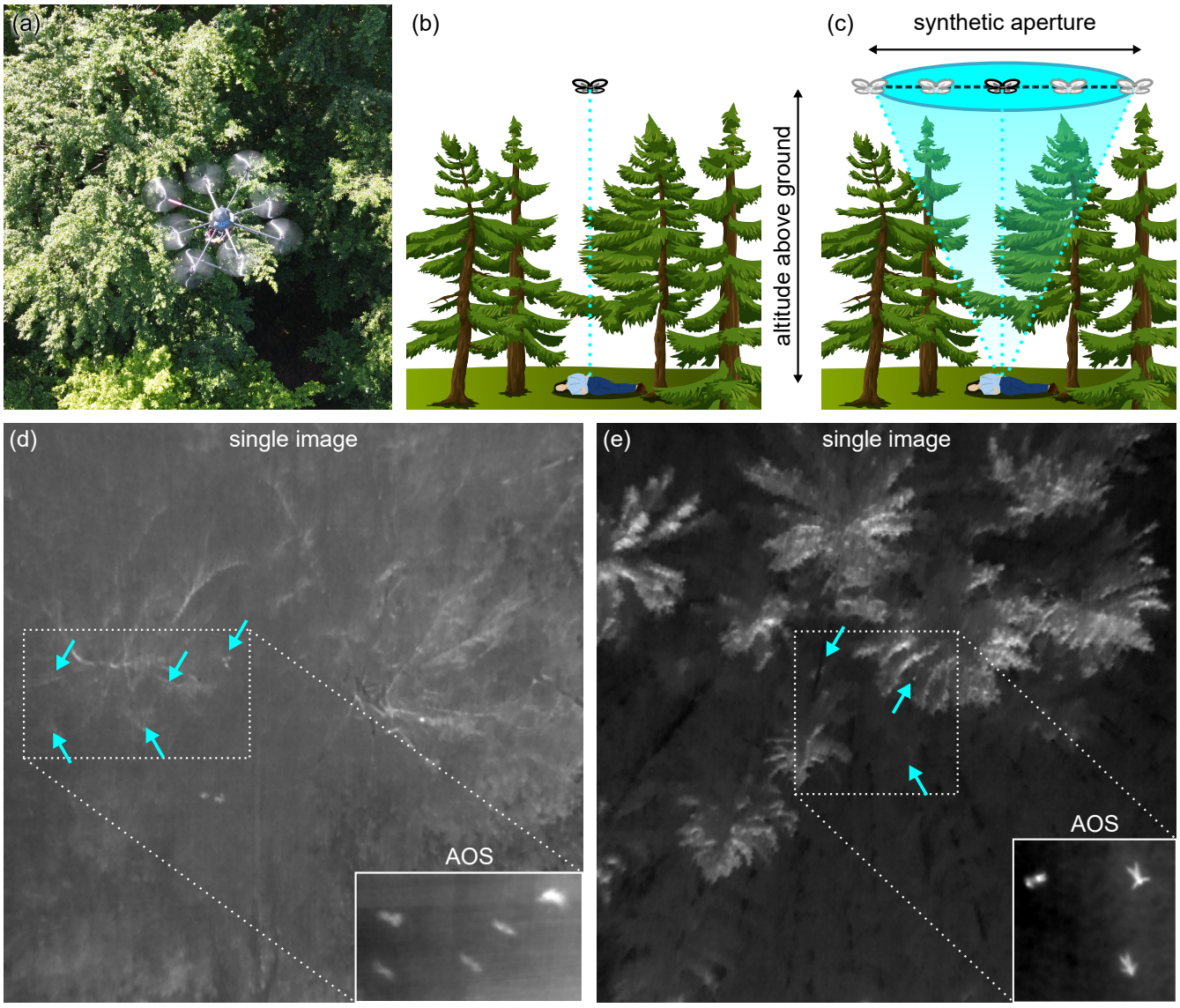

Figure 1. (a) Our drone autonomously scanning a forest patch. In contrast to recording and analyzing single images (b), AOS combines multiple images that are captured within a synthetic aperture before the resulting integral image is analyzed (c). Single thermal drone recordings at an altitude of 35 m above dense forest ground: (d) mixed forest, (e) conifer forest. The arrows indicate partial heat signals of occluded people on the ground. The insets show AOS results that are achieved when multiple thermal images are integrated. Note that contrast and brightness of the insets have been adjusted for better visibility. See Supplementary Video.

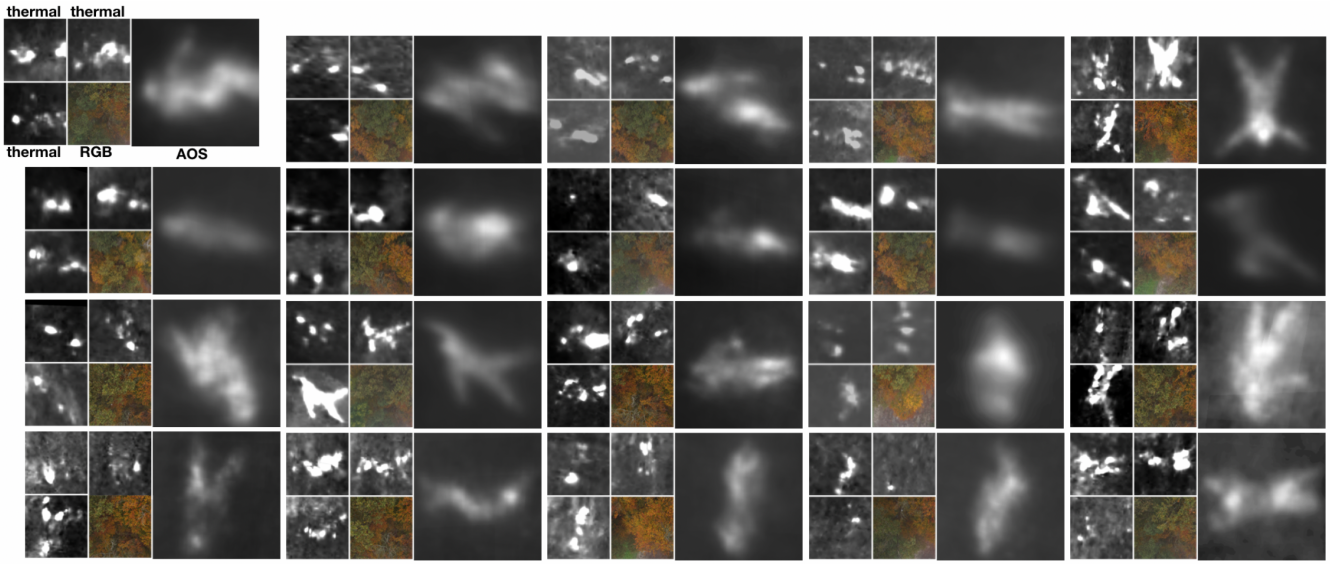

Figure 2. Results of an initial field experiment. Twenty people lying on the ground in a dense broadleaf forest (zoom to the RGB samples for reference). For each person, a subset of 3 single thermal images (close-ups) and the corresponding AOS integral image (close-up) are shown. The 30 m × 30 m synthetic aperture was scanned within 10 min at an altitude of 35 m above ground.

(see Section S2 in the Supplementary Material for an initial evaluation).

Results of our initial field experiments (cf. Fig. 3) show that occlusion has little effect on the AOS image-forming results. The more occlusion, the lower the contrast in AOS integral images, as explained by the statistical model presented in previous work.[28] After contrast adjustment, however, human shapes and optical aberrations (defocus) can be identified, regardless of whether occlusion was present. This indicates that a training dataset can be produced under controlled open-field conditions (i.e., without occlusion) rather than in forests of different types and densities.

Results

Test and training data for our experiments were recorded in 18 drone flights (see Table 1; Section S3 in the Supplementary Material) in close proximity to Linz, Austria. Twelve flights were performed above forests of various vegetation types (broad-leaved, conifer and mixed forests) at an altitude of approximately 30 m to 35 m above ground layer (AGL). The remaining 6 flights recorded data from above a meadow without any high vegetation. To protect subjects in this open field, a safety net (2 m AGL) was installed. The net strings are not resolvable in integral images captured from the recording altitude.

Subjects were asked to lie on the ground (each in a random pose) and remain still or perform little motion (such as waving hands) to simulate ill or injured persons. Flights covered a square synthetic aperture area of 30 m × 30 m with 1 m × 3 m dense sampling (exceptions are indicated in Table 1). A low-resolution (640 px × 512 px) thermal and a high-resolution (6 000 px × 4 000 px) RGB camera were triggered simultaneously while the drone was flying at a speed of 0.7 m s^{-1}, capturing approximately 360 thermal and RGB image pairs in the course of approximately 10 min flights.

After recording, the high-resolution RGB images were used for precise pose estimation. Since the drone's altitude above ground was known, approximate focus parameters could be pre-estimated (for larger terrain, digital elevation models can be used). Minor variations in ground elevation were handled by a local parameter optimization.[32] Thus, occluding vegetation was suppressed and focused humans on the ground were emphasized in thermal integral images, as illustrated in Fig. 1. In each integral image, people were manually labeled by polygonal contours. Since camera poses and focus parameters are known, the polygonal contour points are three-dimensional and can be related to other camera poses or focus settings. The number of labels (persons) for each flight is shown in Table 1.

The recorded data was split into training, validation and test sets. As previously mentioned, our initial experiments verified that human shapes and optical aberrations (defocus) can be identified in integral images—both with and without occlusion. Thus, for training we used 5 open-field scenes with 37 labels, 2 empty forest scenes, and 1 open-field scene with 5 labels for validation.

For classifying people in our data, we trained the You Only Look Once (YOLO)[33–35] deep learning object detector, which supports fast detection rates, can run on embedded low-cost and low-power systems[36–38] (e.g., as used on drones), and has proved its applicability to thermal object detection tasks in previous studies.[39,40]

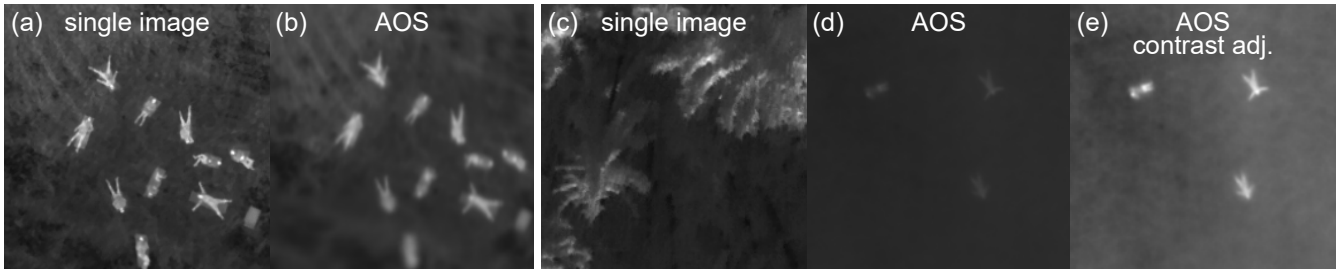

Figure 3. Comparison of scenes without (a–b) and with (c–e) occlusion in terms of contrast level. The differences between AOS integral images without (b) and with (d) occlusion are small after contrast adjustment (e).

Table 1. Experimental dataset (see Fig. S3 for exemplary RGB and thermal images). Latitude angles are north of the equator and longitude angles are east of the prime meridian.

ID	Latitude	Longitude	Forest	Date	Labels	Set
F0[a]	48°25′17.32″	14°18′10.08″	conifer	4 Oct 19	3	test
F1[b]	48°19′59.42″	14°19′52.40″	broadleaf	24 Oct 19	10	test
F2	48°19′59.56″	14°19′52.77″	broadleaf	24 Oct 19	10	test
F3	48°20′01.41″	14°19′48.50″	mixed	25 Oct 19	6	test
F4	48°20′01.56″	14°19′48.74″	mixed	25 Oct 19	6	test
F5	48°19′57.70″	14°19′48.35″	conifer	8 Nov 19	10	test
F6	48°19′57.70″	14°19′48.35″	conifer	8 Nov 19	10	test
F7[c]	48°19′59.56″	14°19′52.18″	broadleaf	20 Nov 19	2	test
O1	48°20′16.46″	14°18′50.52″	none	8 Jan 20	10	train
O2	48°20′16.46″	14°18′50.52″	none	8 Jan 20	10	train
F8	48°19′59.19″	14°19′52.11″	broadleaf	17 Jan 20	0	train
F9	48°19′58.66″	14°19′51.71″	broadleaf	17 Jan 20	0	test
O3	48°20′16.46″	14°18′50.52″	none	22 Jan 20	6	train
O4	48°20′16.46″	14°18′50.52″	none	22 Jan 20	6	train
O5	48°20′16.46″	14°18′50.52″	none	7 Feb 20	5	train
O6	48°20′16.46″	14°18′50.52″	none	7 Feb 20	5	valid
F10	48°20′01.75″	14°19′48.92″	mixed	10 Apr 20	0	train
F11	48°19′57.60″	14°19′48.39″	conifer	10 Apr 20	0	test

[a] 1 m × 2 m spacing, resulting in 402 images.
[b] aborted early and contains only 153 images.
[c] 5 m circular synthetic aperture with 31 images.

To increase the number of samples in the training and validation sets, we used common data augmentation techniques. For instance, we optionally applied adaptive histogram equalization (AHE) to every integral image and added the resulting images to the dataset. AHE reduces the effect of temperature variations and has been used previously to enhance thermal images.[41] The orientation of the images was also changed randomly 10 times to account for various rotations of the test subjects. Furthermore, we altered the focus parameters, because multiple heat sources on the ground or non-planar ground may lead to slight defocus for a single focal plane setting. We applied 27 focus variations by translating focus away and towards the ground and by rotating the focal plane about its two axes.

Since our labels are 3D polygonal contours, they can be converted to correct axis-aligned bounding boxes after augmentation. Ultimately, augmentation resulted in a total of 540 images with corresponding labels for every flight. Further augmentations were performed during training by the training algorithm of YOLO and include random horizontal image flipping and minor brightness changes. We used the average precision (AP) metric[42] for the validation dataset to determine when to stop training.

After training, we applied the resulting network to the test dataset. At test time, we optionally performed AHE, ran the detection twice (on both the non-augmented and the augmented images), and combined the two results. To avoid potential double detections, we applied non-maximum suppression (NMS).

For AHE applied to both training and test set, we achieved an overall AP score of 92.2% (precision/recall 96.4/93%; IoU threshold 25%) and detected 53 out of 57 persons (true positives) and only 2 false positives (branches and a dog, were classified as persons) in the test dataset. Visual results are shown in Fig. 4.

We compared AOS detection to a conventional single-image detection in terms of performance, and classified all single images of the test set. For this purpose, a new network was trained with the single images of the training set. Three-dimensional labels were transferred automatically from the integral images to the single images. The augmentations used for the integral images were reused, with the exception of the focus parameters (which do not apply to single images). After training, the network was applied to the single images of the test set. As indicated by our initial experiments, the single-image detection rate dropped significantly to a maximum AP score of 24.8% (precision/recall scores 25.1/50.4%) when AHE was applied to the test set only (cf. Fig. 5).

Tables 2 and 3 show AP scores, ground truth (GT), and the number of true positives (TP) and false positives (FP) for our test scenes (rows) when using AOS and single images, respectively. The tables list the detection results for the networks trained with and without AHE. For testing, results with and without AHE applied in combination with NMS are listed.

As shown in Table 2, applying AHE to the training and test datasets clearly resulted in the best detection performance for AOS integral images. For single images, AHE reduced FP when applied to the test set (i.e., the average FP rate dropped from 4.5 to 0.7 without and from 5.4 to 1.5 with AHE test augmentation), but also reduced the TP rate by a factor of approximately 2. This drop explains the lower AP scores of the results with training augmentation compared to those without, as summarized in Table 3.

Conclusion and future work

We have shown that the detection rate for human classification under occlusion conditions can be significantly increased by combining multi-perspective recordings before classification. While commonly used real-time classifiers (such as YOLO) produce poor results on single images, they produce usable outputs on integral images. We have also demonstrated that the training data for our approach is widely invariant to occlusion, and can therefore be generated easily under controlled (open-field) conditions.

We believe, that these findings provide a foundation for future autonomous SAR technologies that focus on finding lost and injured people in dense forests. Our approach can also assist conventional SAR missions that are carried out with manned helicopters or airplanes. Furthermore, it could support the surveillance of humans in the course of military and law-enforcement tasks, the monitoring of animals for wildlife observation, or autonomous vehicles whenever person classification suffers from occlusion. However, with a world-wide increase of drone applications, new challenges, concerning ethics, sustainability, security, and privacy, arise and need to be addressed.[43–45] Camera drones should adhere to privacy regulations and respect human rights.

Registration and averaging is just one possible way of combining images from multi-perspective recordings with AOS, and we are planning to explore further options. For instance, AOS also supports computing entire focal stacks (i.e., integral images for multiple, axially varying synthetic focal planes), which better preserve depth information than single integral images and might increase detection rates further. This approach requires a network structure that operates on volumetric data,[46] which we will investigate as part of future work.

Our current implementation could clearly be improved by more sophisticated detection techniques[47–50] and larger training sets with more human participants. Advanced data augmentation techniques might also lead to higher detection rates. First experiments applying augmentation techniques that simulate occlusions in single images, however, have not improved the detection performance and are described in Section S4 of the Supplementary Material. Notwithstanding these considerations, we believe that combining multiple perspectives before classification will continue to produce superior results. One of our biggest limitation is the short battery life of camera drones that restricts flight time and thus the scanning coverage. Therefore, we are currently investigating the efficiency

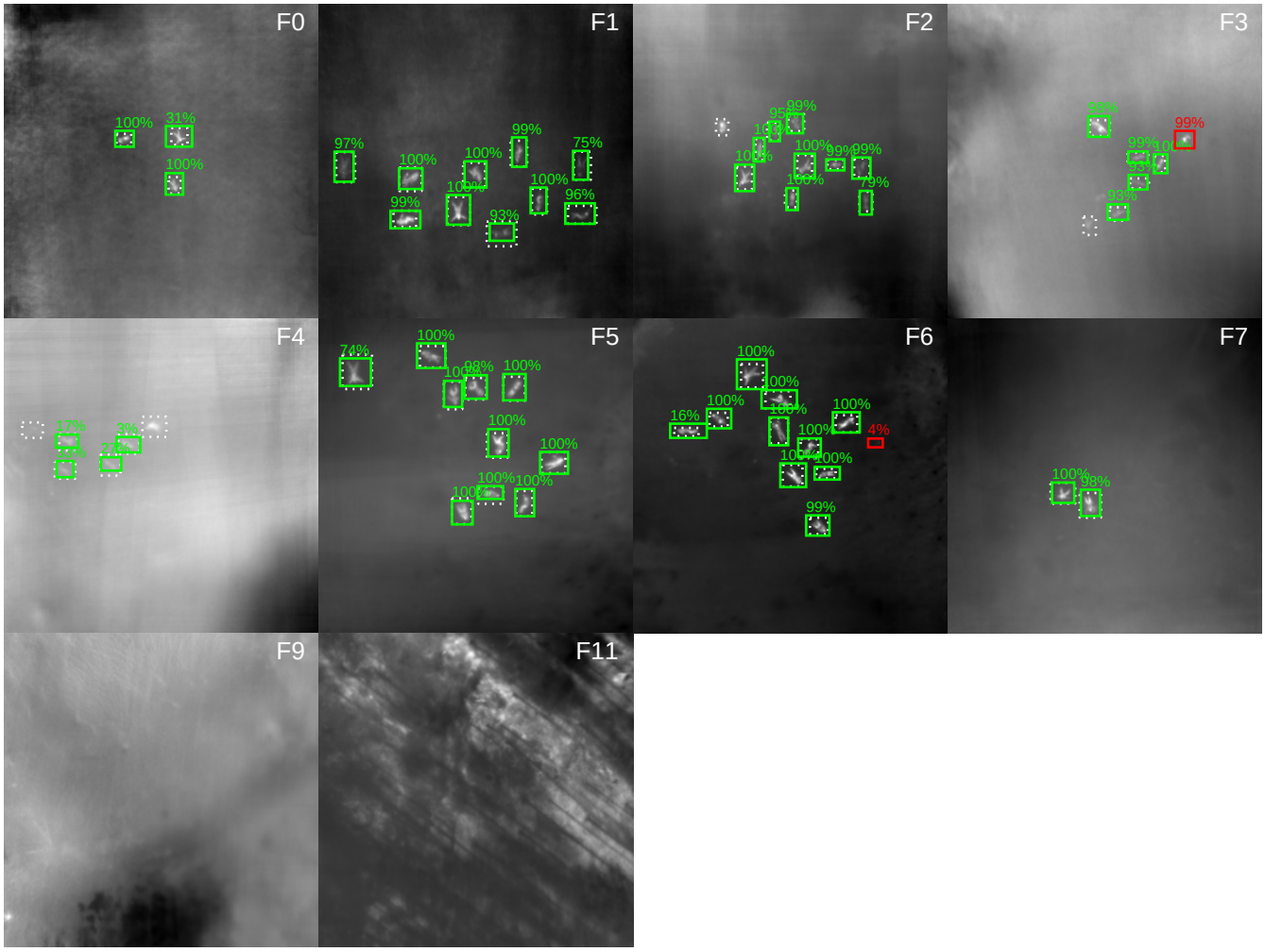

Figure 4. AOS person detection results for the 10 test scenes (cf. Table 1, AHE augmentation applied to both test and training sets). The ground truth labels are enclosed within white dashed rectangles. Detections are indicated by solid rectangles, where TPs are green and FPs are red. The numbers above bounding boxes indicate the network's confidence score. Corresponding AP scores and the numbers of TPs and FPs are shown in Table 2 (column "AHE train + test set"). Note that F9 and F11 are empty.

Table 2. AOS person detection results. Average precision scores (AP), ground truth (GT), true positives (TP), and false positives (FP) for the integral image of each scene. Augmentation with AHE: not applied, applied to test set only, applied to training set only, applied to test and training set.

	no AHE			**AHE test set only**			**AHE train set only**			**AHE train + test set**		
ID (GT)	**AP**	**FP**	**TP**	**AP**	**FP**	**TP**	**AP**	**FP**	**TP**	**AP**	**FP**	**TP**
F0 (3)	**100.0%**	0	**3**	**100.0%**	0	**3**	**100.0%**	0	**3**	**100.0%**	0	**3**
F1 (10)	**100.0%**	0	**10**	**100.0%**	0	**10**	60.0%	0	6	**100.0%**	0	**10**
F2 (10)	80.0%	0	8	**90.0%**	0	**9**	30.0%	0	3	**90.0%**	0	**9**
F3 (6)	16.7%	0	1	50.0%	0	3	16.7%	0	1	**73.1%**	1	**5**
F4 (6)	0.0%	0	0	16.7%	0	1	0.0%	0	0	**66.7%**	0	**4**
F5 (10)	**100.0%**	0	**10**	**100.0%**	0	**10**	90.0%	0	9	**100.0%**	0	**10**
F6 (10)	**100.0%**	1	**10**	99.1%	2	**10**	**100.0%**	0	**10**	**100.0%**	1	**10**
F7 (2)	**100.0%**	0	**2**	**100.0%**	1	**2**	50.0%	0	1	**100.0%**	0	**2**
F9 (0)	n/a	0	0	n/a	0	0	n/a	0	0	n/a	0	0
F11 (0)	n/a	0	0	n/a	0	0	n/a	0	0	n/a	0	0
sum (57)	77.2%	1	44	83.2%	3	48	57.9%	0	33	**92.2%**	2	**53**

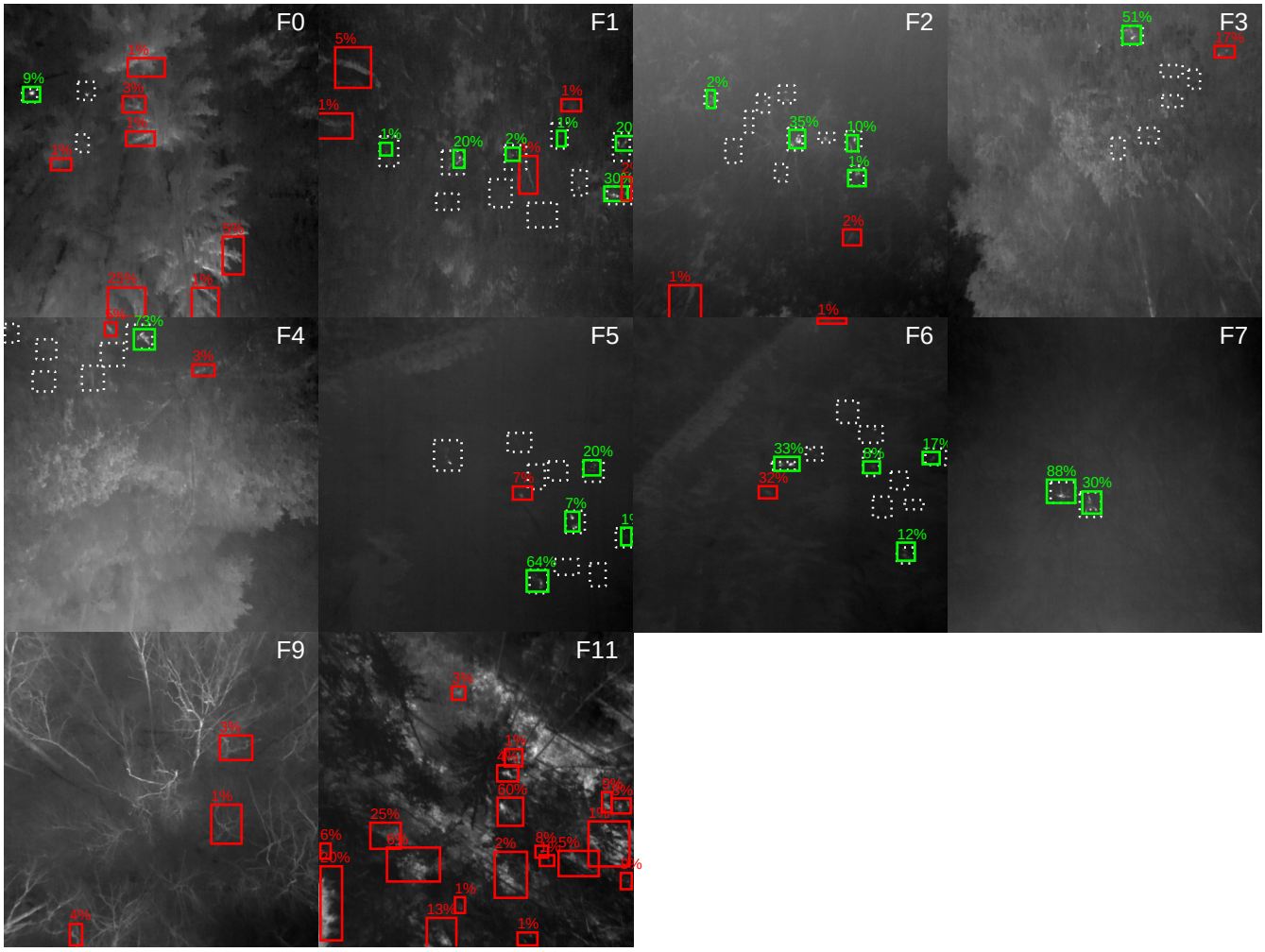

Figure 5. Single-image person detection results for the 10 test scenes (cf. Table 1, AHE augmentations applied to test set only). The ground truth labels are enclosed within dashed rectangles. Detections are indicated by solid rectangles, where TPs are green and FPs are red. The numbers are the network's confidence scores. Corresponding AP scores and the numbers of TPs and FPs are shown in Table 3 (column "AHE test set only").

Table 3. Single-image person detection results. Average precision scores (AP), ground truth (GT), true positives (TP), and false positives (FP) as averages over all images of each scene. Augmentation with AHE: not applied, applied to test set only, applied to training set only, applied to test and training set.

	no AHE			AHE test set only			AHE train set only			AHE train + test set		
ID (avg. GT)	**AP**	**FP**	**TP**	**AP**	**FP**	**TP**	**AP**	**FP**	**TP**	**AP**	**FP**	**TP**
F0 (2.6)	4.0%	7.1	0.8	**7.6%**	8.8	**1.0**	4.3%	1.4	0.4	5.0%	1.6	0.4
F1 (7.2)	55.5%	4.3	4.6	**57.1%**	4.7	**4.7**	27.5%	0.7	2.1	28.1%	0.8	2.1
F2 (8.5)	24.5%	2.6	2.8	**33.7%**	3.4	**3.7**	11.6%	0.1	1.0	11.7%	0.2	1.0
F3 (4.4)	14.7%	0.7	0.8	**20.9%**	1.4	**1.2**	3.2%	0.0	0.2	3.1%	0.1	0.2
F4 (2.6)	3.4%	1.2	0.3	**7.5%**	2.3	**0.5**	1.9%	0.1	0.1	1.8%	0.2	0.1
F5 (5.8)	57.3%	1.0	3.5	**63.5%**	1.2	**3.8**	34.5%	0.3	2.1	35.0%	0.4	2.1
F6 (5.7)	72.6%	2.0	4.3	**75.7%**	2.2	**4.4**	55.2%	0.7	3.2	55.9%	0.9	3.2
F7 (2.0)	75.4%	0.1	1.5	**96.6%**	0.2	**1.9**	51.5%	0.0	1.0	51.5%	0.0	1.0
F9 (0.0)	n/a	2.1	0.0	n/a	2.9	0.0	n/a	1.6	0.0	n/a	4.5	0.0
F11 (0.0)	n/a	16.5	0.0	n/a	18.0	0.0	n/a	0.8	0.0	n/a	4.8	0.0
avg (3.6)	18.0%	4.5	1.6	**24.8%**	5.4	**1.8**	19.0%	0.7	0.8	17.7%	1.6	0.9

of one-dimensional line scans (i.e., 1D synthetic apertures) for person detection, rather than two-dimensional area scans. Initial experiments indicate that 1D apertures are sufficient and allow to cover a significantly larger range. Furthermore, we are working on a first fully embedded on-board implementation that carries out all measurements and computations directly on the drone and during flight. It will support real-time rates (so far, approximately 200 ms are needed for image integration and classification using a Raspberry Pi and an Intel Neural Compute Stick that runs YOLO-tiny[35] (a YOLO version optimized for mobile processors).

Methods

We recorded our datasets using an octocopter (MikroKopter OktoXL 6S12; 945 mm diameter; approx. 4.9 kg; two LiPo 4 500 mA h batteries), equipped with a thermal camera (Flir Vue Pro; 9 mm fixed focal length lens; 7.5 µm to 13.5 µm spectral band; 14 bit non-radiometric) and an RGB camera (Sony Alpha 6000; 16 mm to 50 mm lens at infinite focus). The cameras were fixed to a rotatable gimbal, were triggered synchronously (synched by a MikroKopter CamCtrl control board), and pointed downwards during all flights. A synthetic aperture of 30 m × 30 m was chosen, because at an altitude of 30 m (maximal tree height plus safety margin) the field of view of all recorded single images is just overlapping on the ground. The aperture's flight pattern was planned using MikroKopter's flight planning software and uploaded to the drone as waypoints. The waypoint protocol triggered the cameras every 1 m along the flight path, and the recorded images were stored on the cameras' internal memory cards.

After landing the drone and downloading the images from the memory cards, we processed the recorded data on a personal computer. To estimate the drone's pose, we used the unprocessed RGB images (6 000 px × 4 000 px) together with the general-purpose structure-from-motion and multi-view stereo pipeline, COLMAP.[51] COLMAP required approximately 24 minutes for pose estimations of 300 images in our implementation. Since the cameras were fixed to a gimbal, the poses of the thermal camera could be directly obtained from the poses of the RGB camera by means of a precalibrated transformation matrix, which was computed using Matlab's checkerboard calibration routine. The calibration checkerboard was made of metal with black velvet checkers and is detectable in both thermal and RGB images.[27] The thermal images were rectified to remove lens distortions and cropped to a field of view of 50.82° and a resolution of 512 px × 512 px. For rectification we applied OpenCV's pinhole camera model.[52] Since our thermal camera was non-radiometric, sensor readings did not correspond to absolute (but to relative) temperatures and changed continuously. Therefore, the thermal images' intensity mean was adjusted to the same range for each recorded scene.

The integral images were computed on a GPU using our visualization technique[25] implemented with Nvidia's CUDA toolkit. The integration of 360 single images took 60 ms on our system. For integral image visualization, a virtual camera was placed within the synthetic aperture's center, and its field of view was set to 50.82° (i.e., the single-image field of view after rectification). Optimal settings for the synthetic focal plane were obtained by optimization.[32] Note that the automatic focal plane optimization did not focus on the ground of scenes F8 and F9, as there were no distinguishable heat sources. Thus, focal plane adjustment was done manually in these two cases. More details on generating AOS integral images can be found in our previous publications[25,27] and in Section S1 of the Supplementary Material.

Persons were labeled in single (non-augmented) integral images using Matlab's polygonal tool. The labels were converted to axis-aligned bounding boxes (AABBs) after augmentation (i.e., rotation) and stored as text files at the location of the corresponding images. Although the classifier as well as AP, TP and FP computations require AABBs, we used polygonal labels as intermediate representations, since they remain unaffected by rotation augmentation.

For the test and the validation datasets (see Table 1), we applied the following augmentation techniques to the integral images: We randomly rotated the images by changing the direction of the virtual camera's up-vector in our visualization (corresponding to a rotation about the image center). Furthermore, we modified the focal plane parameters. We changed the altitude of the focal plane by ± 0.25 m around the optimal focal plane, and rotated it about its vertical and horizontal axes (relative to the virtual camera's orientation) by ± 2°. This led to a fixed number of 270 augmentations per scene. Note that our augmentation pipeline operated on single-precision floating-point (32 bit) high-dynamic-range (HDR) images. The optional adaptive histogram equalization was performed using Matlab's Image Processing Toolbox after all other augmentation steps had been applied. For HDR images, the number of histogram bins was increased to 512 while all other parameters were kept at their default settings. Finally, the HDR images were tone-mapped to 8-bit LDR grayscale images (required by the classifier) and normalized within the range of the hottest and coldest relative temperature readings per image.

Augmentations of single images (i.e., rotation and optional AHE) were performed in Matlab, as explained above. Blank borders that were introduced when rotating rectangular images were removed by cropping. The labels of the integral images were directly projected to the single images using the known pose matrices and focal plane parameters (the contours define a 3D plane). Projected labels at borders were truncated and discarded if their clipped AABBs were less than 25% of their previous sizes. After label transfer, all images were inspected manually and outliers (due to, e.g., strong pose estimation errors) were removed.

We trained YOLO[35] version 3 as implemented in the Darknet framework,[53,54] employing the network structure that utilizes spatial pyramid pooling (SPP).[55,56] The default YOLOv3 SPP configuration with 114 layers was used, and only the

changes required to support single-class (person) detection were applied. The input image size of the network was set to 512 px × 512 px and matched the resolution of the integral images and the rectified (non-cropped) single images. Training was performed on two NVIDIA GTX 2070 GPUs, and the network's batch and subdivision sizes were set to 64 to fit into 8 GB of GPU RAM. During training, YOLOv3 performed further augmentations internally, including random horizontal image flipping and minor brightness changes. Augmentations unsuitable for grayscale images (e.g., hue or saturation changes) were turned off. For training we used 53 convolutional filters that were pre-trained on Imagenet[57] for the first 75 network layers (the backbone). The starting learning rate was set to 0.001, and training weights were stored after every 200 batch iterations. After training, the stored training weights were used to compute AP scores (with IoU 50% using Darknet) on the validation datasets, and the weight with the highest AP result was used as final weight. Note that for the validation and training sets the same training augmentations were applied. Test augmentations were not applied to the validation datasets. The optimal weights were obtained after 3 200 to 4 600 iterations or 4 to 143 epochs (cf. Figure S5 and Section S5 of the Supplementary Material).

For evaluation, we applied the trained networks to our test datasets (see Table 1). We ran Darknet on the test images and stored corresponding detections (i.e., bounding-box locations and the confidence score of the network). Predictions for one image were computed in 30 ms in our implementation. Detections with a confidence score below 0.5% were discarded. The test-time augmentation (including NMS) and AP, TP, and FP scores (as reported in Tables 2 and 3) were computed in Matlab. For NMS and for AP, TP and FP scores we used an IoU threshold of 25%. To account for the AABB clipping and to avoid false defections at the image borders, we discarded detection results for which the AABB's center was too close to the image border. Since the median bounding box size in the training set was 35 px, we used a distance threshold of 5 px (half of 35 px×25%, rounded up). Note that this only affects single-image detection results, as our integral images require no AABB clipping. For integral images, we obtained one detection result per scene (see Table 2 and Figure 4). For single images, we averaged detection results over all single images per scene (see Table 3). Figure 5 shows a representative selection of single-image classification results. Precision and recall plots are shown in Figure S6 and discussed in Section S6 of the Supplementary Material.

Ethical approval.

The ethics committee of the Upper Austrian government approved the study, and participants provided written informed consent.

Data availability

The data collected in experiments with users can be downloaded from Zenodo[58], and includes labels and augmented images for training, validation, and testing, configuration files, trained network weights, and results.

Code availability

Code to compute Tables 2 and 3 is provided with the dataset.[58] Further code that supports the findings of this study is available from the corresponding author upon reasonable request.

References

1. Burke, C. *et al.* Requirements and limitations of thermal drones for effective search and rescue in marine and coastal areas. *Drones* **3**, 78 (2019).
2. Lygouras, E. *et al.* Unsupervised Human Detection with an Embedded Vision System on a Fully Autonomous UAV for Search and Rescue Operations. *Sensors* **19**, 3542, DOI: 10.3390/s19163542 (2019).
3. Brunetti, A., Buongiorno, D., Trotta, G. F. & Bevilacqua, V. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing* **300**, 17–33 (2018).
4. Yurtsever, E., Lambert, J., Carballo, A. & Takeda, K. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access* **8**, 58443–58469 (2020).
5. Moreira, A. *et al.* A tutorial on synthetic aperture radar. *IEEE Geosci. Remote. Sens. Mag.* **1**, 6–43, DOI: 10.1109/MGRS.2013.2248301 (2013).
6. Li, C. J. & Ling, H. Synthetic aperture radar imaging using a small consumer drone. In *2015 IEEE International Symposium on Antennas and Propagation USNC/URSI National Radio Science Meeting*, 685–686, DOI: 10.1109/APS.2015.7304729 (2015).
7. Rosen, P. A. *et al.* Synthetic aperture radar interferometry. *Proc. IEEE* **88**, 333–382, DOI: 10.1109/5.838084 (2000).
8. Levanda, R. & Leshem, A. Synthetic aperture radio telescopes. *Signal Process. Mag. IEEE* **27**, 14 – 29, DOI: 10.1109/MSP.2009.934719 (2010).
9. Dravins, D., Lagadec, T. & Nuñez, P. D. Optical aperture synthesis with electronically connected telescopes. *Nat. communications* **6**, 6852, DOI: 10.1038/ncomms7852 (2015).
10. Ralston, T. S., Marks, D. L., Carney, P. S. & Boppart, S. A. Interferometric synthetic aperture microscopy (ISAM). *Nat. Phys.* 965–1004, DOI: doi:10.1038/nphys514 (2007).
11. Hayes, M. P. & Gough, P. T. Synthetic aperture sonar: a review of current status. *IEEE J. Ocean. Eng.* **34**, 207–224 (2009).
12. Hansen, R. E. Introduction to synthetic aperture sonar. In *Sonar Systems Edited* (InTech Published, 2011).

13. Jensen, J. A., Nikolov, S. I., Gammelmark, K. L. & Pedersen, M. H. Synthetic aperture ultrasound imaging. *Ultrasonics* **44**, e5 – e15, DOI: https://doi.org/10.1016/j.ultras.2006.07.017 (2006). In Proceedings of Ultrasonics International (UI'05) and World Congress on Ultrasonics (WCU).

14. Zhang, H. K. *et al.* Synthetic tracked aperture ultrasound imaging: design, simulation, and experimental evaluation. *J. medical imaging (Bellingham, Wash.)* **3**, 027001–027001 (2016).

15. Barber, Z. W. & Dahl, J. R. Synthetic aperture ladar imaging demonstrations and information at very low return levels. *Appl. optics* **53**, 5531–5537, DOI: 10.1364/AO.53.005531 (2014).

16. Turbide, S., Marchese, L., Terroux, M. & Bergeron, A. Synthetic aperture lidar as a future tool for earth observation. *Proc.SPIE* **10563**, 10563 – 10563 – 8, DOI: 10.1117/12.2304256 (2017).

17. Vaish, V., Wilburn, B., Joshi, N. & Levoy, M. Using plane + parallax for calibrating dense camera arrays. In *In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1, I–I, DOI: 10.1109/CVPR.2004.1315006 (2004).

18. Vaish, V., Levoy, M., Szeliski, R. & and, C. L. Z. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 2331–2338, DOI: 10.1109/CVPR.2006.244 (2006).

19. Zhang, H., Jin, X. & Dai, Q. Synthetic aperture based on plenoptic camera for seeing through occlusions. In Hong, R., Cheng, W.-H., Yamasaki, T., Wang, M. & Ngo, C.-W. (eds.) *In Proceedings of Advances in Multimedia Information Processing – PCM 2018*, 158–167 (Springer International Publishing, Cham, 2018).

20. Yang, T. *et al.* Kinect based real-time synthetic aperture imaging through occlusion. *Multimed. Tools Appl.* **75**, 6925–6943, DOI: 10.1007/s11042-015-2618-1 (2016).

21. Joshi, N., Avidan, S., Matusik, W. & Kriegman, D. J. Synthetic aperture tracking: Tracking through occlusions. In *2007 IEEE 11th International Conference on Computer Vision*, 1–8 (2007).

22. Pei, Z. *et al.* Occluded-object 3d reconstruction using camera array synthetic aperture imaging. *Sensors* **19**, 607 (2019).

23. Yang, T. *et al.* All-in-focus synthetic aperture imaging. In Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T. (eds.) *Computer Vision – ECCV 2014*, 1–15 (Springer International Publishing, Cham, 2014).

24. Pei, Z., Zhang, Y., Chen, X. & Yang, Y.-H. Synthetic aperture imaging using pixel labeling via energy minimization. *Pattern Recognit.* **46**, 174–187 (2013).

25. Kurmi, I., Schedl, D. C. & Bimber, O. Airborne optical sectioning. *J. Imaging* **4**, DOI: 10.3390/jimaging4080102 (2018).

26. Bimber, O., Kurmi, I., Schedl, D. C. & Potel, M. Synthetic aperture imaging with drones. *IEEE Comput. Graph. Appl.* **39**, 8–15, DOI: 10.1109/MCG.2019.2896024 (2019).

27. Kurmi, I., Schedl, D. C. & Bimber, O. Thermal airborne optical sectioning. *Remote. Sens.* **11**, DOI: 10.3390/rs11141668 (2019).

28. Kurmi, I., Schedl, D. C. & Bimber, O. A statistical view on synthetic aperture imaging for occlusion removal. *IEEE Sensors J.* 1–1, DOI: 10.1109/JSEN.2019.2922731 (2019).

29. Schedl, D. C., Kurmi, I. & Bimber, O. Airborne optical sectioning for nesting observation. *Sci. Reports* **10**, 1–7 (2020).

30. Hwang, S., Park, J., Kim, N., Choi, Y. & Kweon, I. S. Multispectral pedestrian detection: Benchmark dataset and baselines. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).

31. Xu, Z., Zhuang, J., Liu, Q., Zhou, J. & Peng, S. Benchmarking a large-scale FIR dataset for on-road pedestrian detection. *Infrared Phys. & Technol.* **96**, 199–208, DOI: https://doi.org/10.1016/j.infrared.2018.11.007 (2019).

32. Kurmi, I., Schedl, D. C. & Bimber, O. Fast automatic visibility optimization for thermal synthetic aperture visualization. *IEEE Geosci. Remote. Sens. Lett.* 1–5 (2020).

33. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788 (2016).

34. Redmon, J. & Farhadi, A. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263–7271 (2017).

35. Redmon, J. & Farhadi, A. Yolov3: An incremental improvement. *arXiv:1804.02767* (2018).

36. Shafiee, M. J., Chywl, B., Li, F. & Wong, A. Fast yolo: A fast you only look once system for real-time embedded object detection in video. *arXiv preprint arXiv:1709.05943* (2017).

37. Vandersteegen, M., Vanbeeck, K. & goedeme, T. Super accurate low latency object detection on a surveillance UAV. *arXiv:1904.02024 [cs]* (2019). ArXiv: 1904.02024.

38. Yang, Y., Guo, B., Li, C. & Zhi, Y. An Improved YOLOv3 Algorithm for Pedestrian Detection on UAV Imagery. In Pan, J.-S., Lin, J. C.-W., Liang, Y. & Chu, S.-C. (eds.)

Genetic and Evolutionary Computing, 253–261 (Springer Singapore, Singapore, 2020).

39. Vandersteegen, M., Van Beeck, K. & Goedemé, T. Real-Time Multispectral Pedestrian Detection with a Single-Pass Deep Neural Network. In Campilho, A., Karray, F. & ter Haar Romeny, B. (eds.) *Image Analysis and Recognition*, vol. 10882, 419–426, DOI: 10.1007/978-3-319-93000-8_47 (Springer International Publishing, Cham, 2018). Series Title: Lecture Notes in Computer Science.

40. Ivašić-Kos, M., Krišto, M. & Pobar, M. Human detection in thermal imaging using yolo. In *Proceedings of the 2019 5th International Conference on Computer and Technology Applications*, 20–24 (2019).

41. Zheng, Y., Izzat, I. H. & Ziaee, S. Gfd-ssd: Gated fusion double ssd for multispectral pedestrian detection. *arXiv preprint arXiv:1903.06999* (2019).

42. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. & Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int J Comput. Vis* **88**, 303–338, DOI: 10.1007/s11263-009-0275-4 (2010).

43. Finn, R. L. & Wright, D. Unmanned aircraft systems: Surveillance, ethics and privacy in civil applications. *Comput. Law & Secur. Rev.* **28**, 184 – 194, DOI: https://doi.org/10.1016/j.clsr.2012.01.005 (2012).

44. Rao, B., Gopi, A. G. & Maione, R. The societal impact of commercial drones. *Technol. Soc.* **45**, 83 – 90, DOI: https://doi.org/10.1016/j.techsoc.2016.02.009 (2016).

45. Shakhatreh, H. *et al.* Unmanned aerial vehicles (uavs): A survey on civil applications and key research challenges. *IEEE Access* **7**, 48572–48634, DOI: 10.1109/ACCESS.2019.2909530 (2019).

46. Lu, H., Wang, H., Zhang, Q., Yoon, S. W. & Won, D. A 3D Convolutional Neural Network for Volumetric Image Semantic Segmentation. *Procedia Manuf.* **39**, 422 – 428, DOI: https://doi.org/10.1016/j.promfg.2020.01.386 (2019).

47. Tan, M., Pang, R. & Le, Q. V. EfficientDet: Scalable and Efficient Object Detection. *arXiv:1911.09070 [cs, eess]* (2020). ArXiv: 1911.09070.

48. Zhang, S., Chi, C., Yao, Y., Lei, Z. & Li, S. Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *CVPR* (2020).

49. Songtao Liu, D. H. & Wang, Y. Learning spatial fusion for single-shot object detection. *arxiv preprint arXiv:1911.09516* (2019).

50. Lee, Y. & Park, J. Centermask: Real-time anchor-free instance segmentation. *CVPR* (2020).

51. Schönberger, J. L. & Frahm, J. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4104–4113, DOI: 10.1109/CVPR.2016.445 (2016).

52. Zhang, Z. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis machine intelligence* **22**, 1330–1334 (2000).

53. Bochkovskiy, A. *et al.* Github: Yolo v3, DOI: 10.5281/zenodo.3693999 (2020).

54. Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. Yolov4: Optimal speed and accuracy of object detection. *arXiv:2004.10934* (2020).

55. He, K., Zhang, X., Ren, S. & Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis machine intelligence* **37**, 1904–1916 (2015).

56. Huang, Z. *et al.* Dc-spp-yolo: dense connection and spatial pyramid pooling based yolo for object detection. *Inf. Sci.* (2020).

57. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis. (IJCV)* **115**, 211–252, DOI: 10.1007/s11263-015-0816-y (2015).

58. Schedl, D. C., Kurmi, I. & Bimber, O. Data: Search and rescue with airborne optical sectioning. https://doi.org/10.5281/zenodo.3894773, DOI: 10.5281/zenodo.3894773 (2020).

Corresponding author

Communication and requests for material should be addressed to Oliver Bimber (email: oliver.bimber@jku.at; orcid: 0000-0001-9009-7827).

Acknowledgements

This research was funded by the Austrian Science Fund (FWF) under grant number P 32185-NBL, and by the State of Upper Austria and the Austrian Federal Ministry of Education, Science and Research via the LIT – Linz Institute of Technology under grant number LIT-2019-8-SEE-114.

Author contributions statement

D.S. and O.B. conceived and designed the experiments. D.S. and I.K. performed the experiments. D.S. and O.B. analyzed the data. D.S. and I.K. contributed materials/analysis tools. D.S. and O.B. wrote the paper.

Competing Interests

The authors declare that they have no competing interests.

Supplementary Material:
Search and Rescue with Airborne Optical Sectioning

David C. Schedl*, Indrajit Kurmi*, and Oliver Bimber*

*first.lastname@jku.at; Johannes Kepler University, Faculty of Engineering and Natural Sciences, Linz, 4040, Austria

S1 Computing AOS Integral Images

In this section we briefly revisit the principles of Airborne Optical Sectioning (AOS)[1–5]. The theoretical basis for AOS are unstructured light fields, which represent a 4D subset of the plenoptic function. The interested reader is referred to surveys[6,7] which cover the topic thoroughly. For AOS, thermal radiation (recorded by a thermal camera) is described by rays within an occlusion volume (forest), as show in Figure S1. We parameterize these rays by their intersection with two parallel planes: the synthetic aperture plane (2D directional coordinate at the drone's altitude above the ground h) and the synthetic focal plane (2D spatial coordinate on the ground). In practise, the drone's positions are not precisely on the synthetic aperture plane and the camera's orientation is not stable in flight. Thus, precise 3D pose estimation (extrinsics) and camera calibration (intrinsics) are necessary (see Methods section of the main article). Once the intrinsic and extrinsic parameters are estimated, a single sensor reading (i.e., a thermal pixel) can be parmeterized as a single ray within the 4D light field, resulting in approximately 94 M rays in total for each recorded scene.

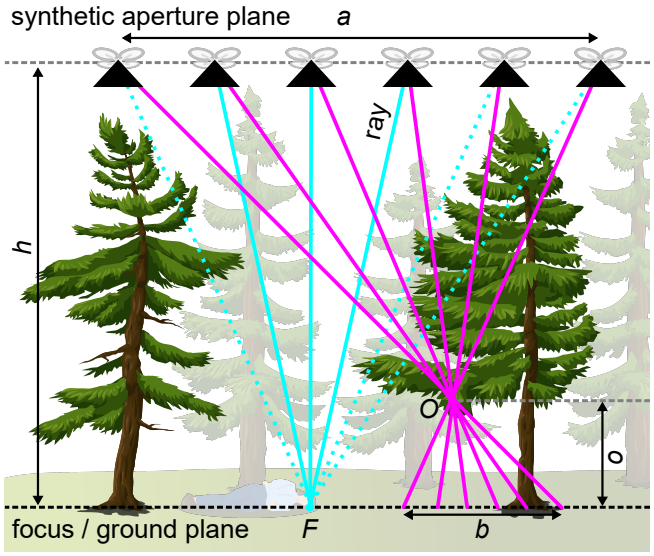

Figure S1. Schematic drawing of AOS' integration principle. Multiple images at the synthetic-aperture-plane of size a and altitude h above ground level are recorded and integrated as a 4D light field. Exemplary rays are shown for a target point F (cyan) on the synthetic focal plane (aligned with the ground surface), and an out-of-focus occluder O (magenta). While the rays of F form a point on the synthetic focal plane, the projection of out-of-focus rays form an area of size b (cf. Equation S1). Dotted in-focus rays indicate occlusion (e.g., a $D = 50\%$ occlusion by vegetation).

For integration, rays that intersect at a given spatial coordinate (i.e., varying directional coordinates on the synthetic aperture plane and similar spatial coordinates on the synthetic focal plane) are averaged (cf. Figure S1 in-focus point F). The result of integrating all rays across all spatial coordinates on the synthetic focal plane is a focused (on the ground) integral image. Setting the focus parameters (i.e., aligning the synthetic focal plane with the unknown ground surface) is done by automatic optimization[8] (see Results and Methods sections of the main article). Note, that changing the focus parameters requires a reparameterization[9] of the 4D ray space, as the spatial coordinates are changing.

In case of occlusion, rays will be D likely to contain signals of occluders, where D is the density of occluders[4]. Occluded rays reduce the contrast (cf. Figure 3) and form a scattered footprint in the integral image (cf. Figure S1 point O). While occluders are blurred and deemphasized, targets on the focused ground (such as humans) can be made clearly visible. The footprint size b of a hypothetical infinitely-small out-of-focus point can be expressed with the intercept theorem as

$$b = \frac{oa}{h-o}\,, \tag{S1}$$

where o is the altitude (above ground level) of the occluder, a is the synthetic aperture size, and h is the height of the synthetic aperture plane above ground level. This means, for example, that an occluder at $o = 2\,\text{m}$ above ground has a footprint $b = 1.8\,\text{m}$, in our experiments ($a = 30\,\text{m}$; $h = 35\,\text{m}$). By considering the occluder size w, Equation S1 extends to

$$b' = b + \frac{wh}{h-o} = \frac{oa+wh}{h-o}\,, \tag{S2}$$

where $(wh)/(h-o)$ is the occluder's projection onto the focal (ground) plane.

S2 Using Existing Image Datasets

In this section, we apply the FLIR advanced driver-assistance systems (ADAS)[10] dataset (used for autonomous driving) for training and evaluate its performance when applied to our single image recordings. The dataset consists of 14 452 annotated thermal images with 50 116 person labels and was recorded with a thermal sensor that is similar to ours (14 bit; 640 px × 512 px). It contains mainly upright standing persons,

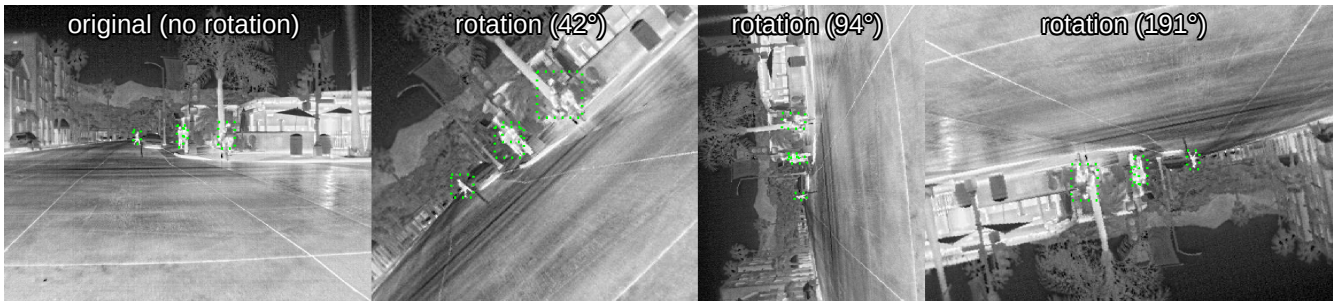

Figure S2. An image from the ADAS dataset used for experiments in Section S2. Exemplary rotations (and crops to avoid the introduction of borders), used for augmentation, are shown. Person labels are indicated by dashed green rectangles. Note that the axis-aligned bounding boxes change size due to rotation.

thus we optionally apply 10 random rotations, in a 360 degree range, on the images and on the labels (cf. Figure S2). Note that we did not apply the optional adaptive histogram equalization (AHE) on the training set for this experiment, as it did not improve single-image results (cf. Table 3 of the main manuscript). We use 90% of the thermal images for training and 10% for validation and perform training as discussed in the Methods section of the main article. The training weights with the highest AP score, when applied to the validation dataset, are obtained after 2400 iterations or 1 epoch. Evaluation results without and with AHE applied to the test images are shown in Table S1. Detection performance on our single images is inferior (the best overall AP scores are below 0.1%), thus indicating that available pedestrian datasets cannot be applied for aerial imaging. The labels in the ADAS datasets contain standing, walking, running and biking persons, while our aerial images show only lying humans. Nevertheless, we believe that the main reason for the poor results is caused by the difference in environments: the ADAS dataset contains persons in hot urban environments, while our recordings show comparatively cool forests with only a few heat spots. Despite the fact that other augmentation techniques (e.g., inverting the temperature scale) may improve detection performance for existing datasets, we believe that our specialized dataset will continue to outperform others. Furthermore, for AOS we rely on known camera poses which are not available for other existing datasets.

S3 Test and Training Sites

Figure S3 shows a representative selection of single RGB and thermal images of all 18 flights at 6 different sites over 10 different days used for the training, validation and test sets. Note that the RGB images are only used for pose estimation (see Methods section in the main article). Scenes labeled with the letter **O** are recorded above our open field training area. The green safety net is clearly visible in the RGB images, but not resolved in the thermal recordings due to the thin strings (4.75 mm) of the net and the comparatively low resolution (640 px × 512 px raw; 512 px × 512 px rectified and cropped) of the thermal camera. Scenes labeled with **F** are recorded

Table S1. Single-image person detection results with a network trained on a pedestrian dataset. Average precision scores (AP), ground truth (GT), true positives (TP), and false positives (FP) for each scene. We compare cases where adaptive histogram equalization (AHE) is applied and not applied to the test set. The scores are clearly worse when compared to Table 3 in the main article.

	no AHE			**AHE test set only**		
ID (GT)	**AP**	**FP**	**TP**	**AP**	**FP**	**TP**
F0 (2.6)	0.01%	0.10	0.00	**0.66%**	16.14	**0.38**
F1 (7.2)	0.00%	0.29	0.00	**0.04%**	2.06	**0.06**
F2 (8.5)	0.00%	0.16	0.00	**0.00%**	3.49	**0.02**
F3 (4.4)	0.04%	0.09	0.01	**0.29%**	8.08	**0.24**
F4 (2.6)	0.00%	0.07	0.00	**0.02%**	7.22	**0.05**
F5 (5.8)	0.00%	0.00	0.00	**0.06%**	0.01	**0.01**
F6 (5.7)	0.00%	0.01	0.00	0.00%	0.02	0.00
F7 (2.0)	0.00%	0.00	0.00	**1.61%**	0.00	**0.03**
F9 (0.0)	n/a	0.02	0.00	n/a	1.54	0.00
F11 (0.0)	n/a	0.62	0.00	n/a	8.40	0.00
avg (3.6)	0.00%	0.15	0.00	**0.09%**	5.80	**0.09**

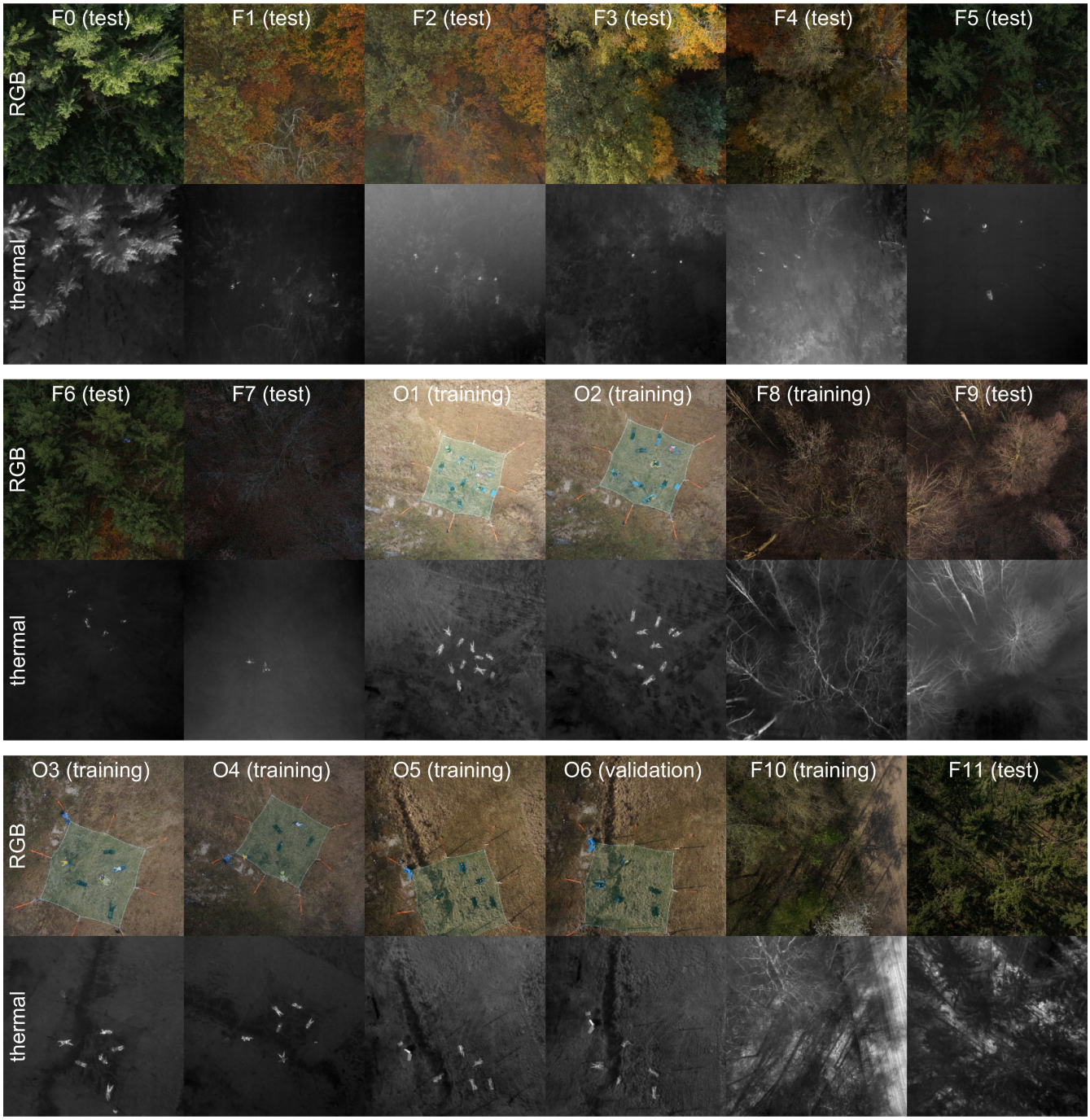

Figure S3. Exemplary RGB and thermal images of the 18 flights that were used to create the training, validation, and test sets. Further details can be found in Table 1 in the main article.

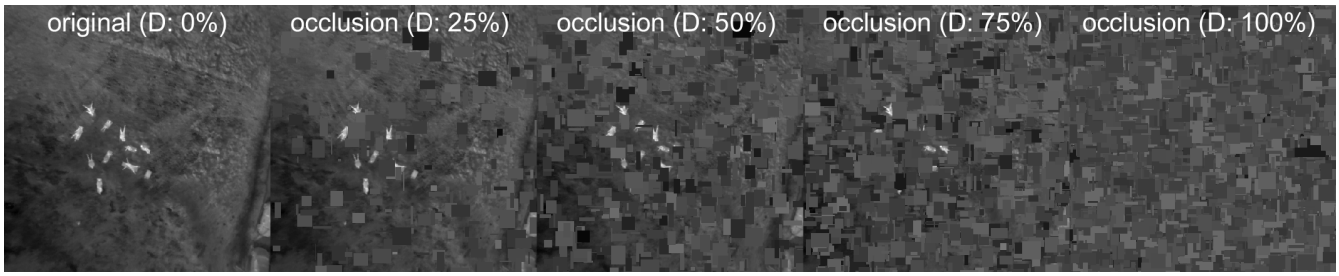

Figure S4. Simulated occlusion for various densities D applied to a single image of the open field training data set.

above forest. The RGB images illustrate the density and vegetation of the different sites. The complete dataset, and flight logs are available open access[11].

S4 Augmenting Simulated Occlusion

In this section we report on initial experiments that augment simulated occlusion to investigate if this can improve the single image detection rate. We apply an augmentation strategy that is similar to random erasing[12] for modelling a Bernoulli distributed occlusion pattern[4]. A random density D (0% to 100%) defines the probability of occlusion in an image (cf. Figure S4). We model occluders as axis-aligned rectangles of random width and height (1 px to 35 px i.e., the average bounding box size in the training set). Thermal values for the occluding rectangles are random samples from non-labeled (i.e., background) regions of the original single images. We apply occlusion augmentation for (initially every occlusion free) image in the training and validation dataset and add it to the corresponding set prior to training. The best training weights are retrieved after 2800 iterations or 4 epochs. Note, that we did not apply the optional adaptive histogram equalization (AHE) to the training set for this experiment, as it did not improve single-image results (cf. Table 3 of the main manuscript). Evaluation results (shown in Table S2), indicate that the detection rate with single images is not improved by augmenting simulated occlusion.

S5 Training Weights

Training weights are stored after every 200 batch iterations and the weight with the highest AP scores (with IoU 50% using Darknet) on the validation datasets are used as final weight. Note, that for the validation and training sets the same training augmentations were applied. Figure S5 plots AP scores on the validation dataset during training of the AOS integral images and the single images. Highest AP scores are achieved after 4200 and 4600 iterations, and 143 and 78 epochs for AOS without and with AHE training augmentation, respectively. For single images, the highest AP scores are achieved after 3600 and 3200 iterations, and 9 and 4 epochs without and with AHE training augmentation. Note, that the number of iterations for one epoch is depending on the number of training images (i.e., more training images requires a higher number of iterations for one epoch). No test augmentations were applied on the validation datasets.

Table S2. Single-image person detection results with simulated occlusion augmentation. Average precision scores (AP), ground truth (GT), true positives (TP), and false positives (FP) for each scene. Augmentation with AHE: not applied, and applied to the test set only. Note, that the scores did not improve, when compared to Table 3 in the main article.

	no AHE			AHE test set only		
ID (GT)	**AP**	**FP**	**TP**	**AP**	**FP**	**TP**
F0 (2.6)	1.9%	7.3	0.5	**2.9%**	7.5	**0.6**
F1 (7.2)	41.9%	1.6	3.4	**44.8%**	1.8	**3.6**
F2 (8.5)	20.3%	1.4	2.2	**22.8%**	1.5	**2.4**
F3 (4.4)	7.0%	0.3	0.4	**10.2%**	0.4	**0.6**
F4 (2.6)	3.0%	0.7	0.2	**4.2%**	0.7	**0.3**
F5 (5.8)	47.1%	0.4	2.8	**53.0%**	0.4	**3.1**
F6 (5.7)	58.3%	0.4	3.4	**60.6%**	0.4	**3.5**
F7 (2.0)	25.3%	0.0	0.5	**87.1%**	0.0	**1.7**
F9 (0.0)	n/a	0.9	0.0	n/a	0.9	0.0
F11 (0.0)	n/a	14.4	0.0	n/a	14.4	0.0
avg (3.6)	13.5%	3.5	1.2	**18.38%**	3.5	**1.3**

S6 Precision Recall Plots

Precision-recall plots for the evaluation results presented in the main article (Table 2 and 3) are shown in Figure S6. We used Matlab to compute the scores and plot the curves. The precision-recall curves are the basis for AP score computations. For AOS (cf. Figure S6(a)) the precision scores are only slightly below 1.0 (due to the low number of FPs), while still achieving maximum recall scores of 0.77, 0.84, 0.58, and 0.93 for no AHE augmentation, test-time AHE, AHE applied to training only, and AHE applied to training and test, respectively. Precision and recall scores for single images (cf. Figure S6(b)) are clearly worse, when compared to AOS results. For all four cases, recall is never higher than 0.5. Ignoring the initial precision scores of 1, the highest single-image precision/recall scores are 0.48/0.44, 0.58/0.50, 0.9/0.24, and 0.87/0.24 for no AHE augmentation, and AHE applied to test only, training only, and to the training and test

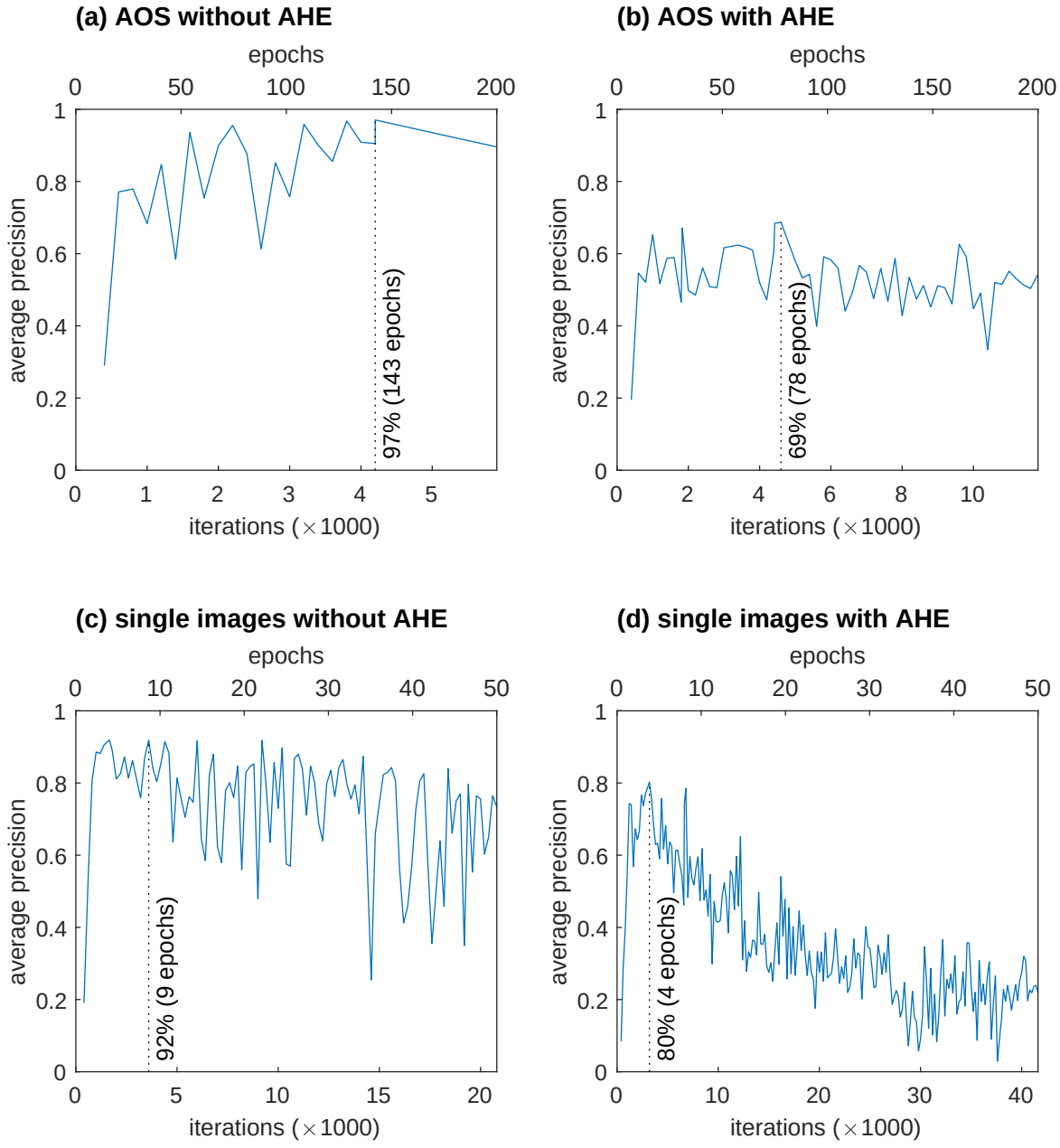

Figure S5. Average precision scores (IoU 50%) on the validation dataset during training for the AOS integral images (a,b) and the single images (c,d), without (a,c) and with (b,d) the optional AHE contrast augmentation. Network weights are evaluated after every 200 training iterations and the weights with maximum AP are used for the evaluation. The bottom x-axis denotes iterations (in thousands) and the top x-axis illustrates the number of epochs. Note that the number of iterations for one epoch is depending on the dataset size (i.e., the number of training images).

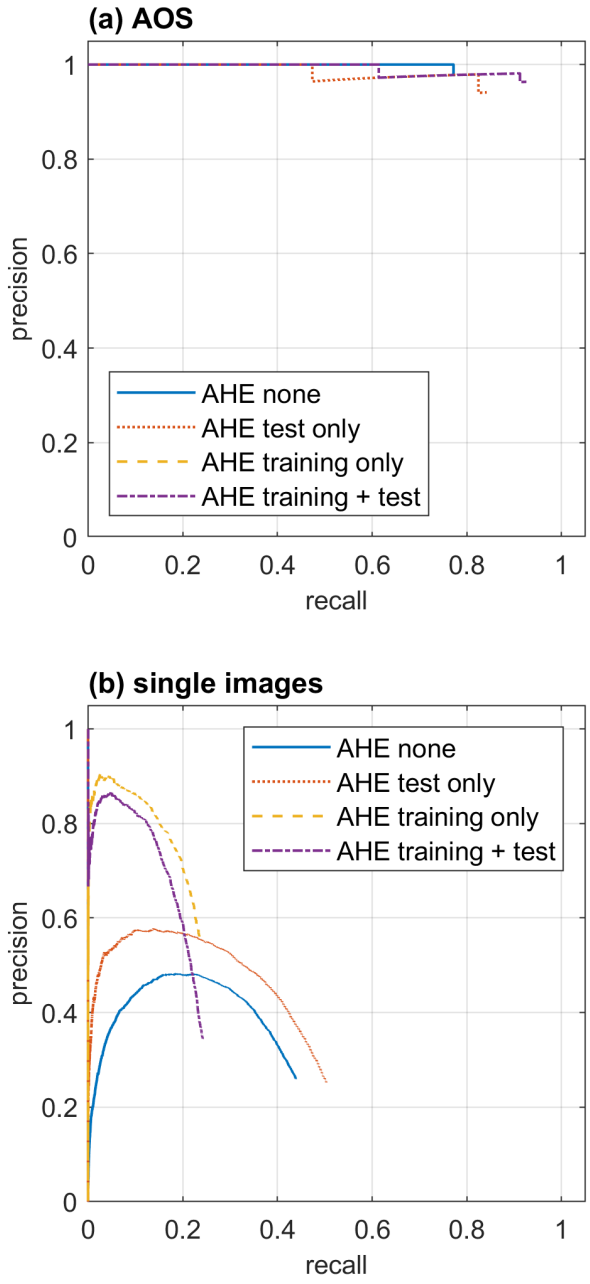

Figure S6. Precision-recall plots for the results presented in the main article (Table 2, (a) and 3, (b)). The blue solid lines indicate results without the optional AHE contrast augmentation. The red dotted lines show results with AHE applied to the test set only, while the orange dashed lines show results with AHE applied to the training set only. The violet, dot and dashed lines show results when AHE is applied to the training and the test set.

set, respectively. The network which is trained without AHE augmentation (AHE none and AHE test only) achieves higher recall but lower precision scores (due to a high number of FPs) when compared to the network that is trained with AHE augmentation on the training set (AHE training only and AHE trainind + test).

References

1. Kurmi, I., Schedl, D. C. & Bimber, O. Airborne optical sectioning. *J. Imaging* **4**, DOI: 10.3390/jimaging4080102 (2018).
2. Bimber, O., Kurmi, I., Schedl, D. C. & Potel, M. Synthetic aperture imaging with drones. *IEEE Comput. Graph. Appl.* **39**, 8–15, DOI: 10.1109/MCG.2019.2896024 (2019).
3. Kurmi, I., Schedl, D. C. & Bimber, O. Thermal airborne optical sectioning. *Remote. Sens.* **11**, DOI: 10.3390/rs11141668 (2019).
4. Kurmi, I., Schedl, D. C. & Bimber, O. A statistical view on synthetic aperture imaging for occlusion removal. *IEEE Sensors J.* 1–1, DOI: 10.1109/JSEN.2019.2922731 (2019).
5. Schedl, D. C., Kurmi, I. & Bimber, O. Airborne optical sectioning for nesting observation. *Sci. Reports* **10**, 1–7 (2020).
6. Wetzstein, G., Ihrke, I., Lanman, D. & Heidrich, W. Computational plenoptic imaging. *Comput. Graph. Forum* **30**, 2397–2426, DOI: 10.1111/j.1467-8659.2011.02073.x (2011). https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2011.02073.x.
7. Wu, G. *et al.* Light field image processing: An overview. *IEEE J. Sel. Top. Signal Process.* **11**, 926–954 (2017).
8. Kurmi, I., Schedl, D. C. & Bimber, O. Fast automatic visibility optimization for thermal synthetic aperture visualization. *IEEE Geosci. Remote. Sens. Lett.* 1–5 (2020).
9. Isaksen, A., McMillan, L. & Gortler, S. J. Dynamically reparameterized light fields. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '00, 297–306 (ACM Press/Addison-Wesley Publishing Co.,, 2000).
10. FLIR thermal starter dataset. https://www.flir.com/oem/adas/adas-dataset-form/ (2018).
11. Schedl, D. C., Kurmi, I. & Bimber, O. Data: Search and rescue with airborne optical sectioning. https://doi.org/10.5281/zenodo.3894773, DOI: 10.5281/zenodo.3894773 (2020).
12. Zhong, Z., Zheng, L., Kang, G., Li, S. & Yang, Y. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (2020).