

A Missing Person Detection in Autonomous SAR System

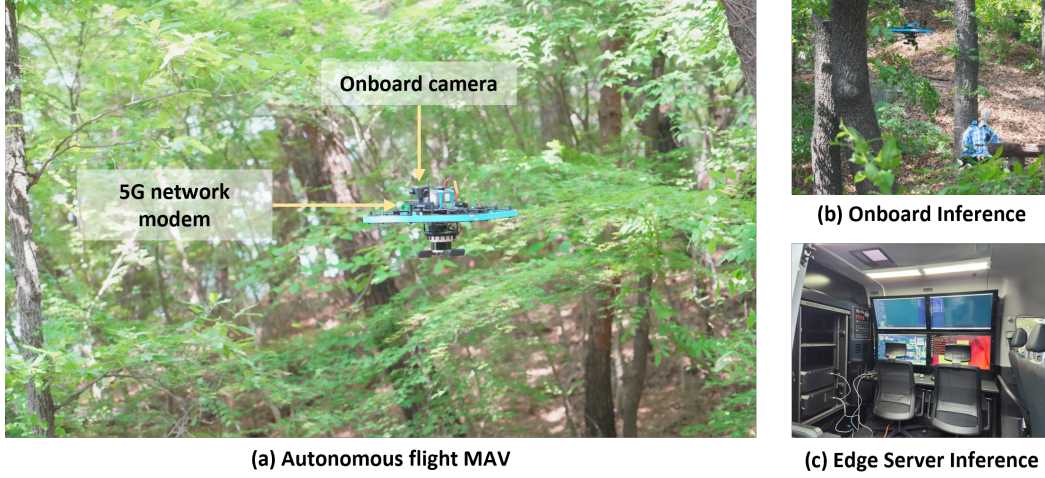


Figure 7: **Missing person detection inference in autonomous SAR systems** (a) Autonomous Flight: A MAV performs a search mission under the forest canopy. (b) Onboard Inference: Frames captured by the onboard camera are processed locally on the MAV in real time using a lightweight detection model. (c) Edge Server Inference: Frames are transmitted via a commercial 5G network to a remote edge server, where inference is performed using a higher-capacity model.

As a future direction and an ongoing application of ForestPersons, we configured a missing person detection pipeline for autonomous Search and Rescue (SAR) missions. In this setup, frames captured by the onboard camera of a Micro Aerial Vehicle (MAV) flying under forest canopy conditions are processed by detection models trained on ForestPersons, which are deployed either onboard the MAV or on a remote edge server depending on mission requirements.

These two inference paths are selected based on the trade-off between latency, bandwidth, and model complexity. For onboard inference, lightweight object detection models such as variants of YOLO [29] are optimized using tools such as NVIDIA TensorRT or Intel OpenVINO to meet real-time constraints on resource-limited hardware, which is particularly useful when low-latency response and independence from network connectivity are critical. In contrast, edge inference allows the use of more advanced models such as transformer-based state-of-the-art architectures like DINO [38]. In this case, video streams are transmitted over a high-bandwidth wireless communication system, such as 5G, to a remote server with greater computational resources. This enables the use of advanced detection algorithms that leverage greater computational resources to achieve improved performance compared to what is feasible on resource-constrained onboard systems.

Figure 7 illustrates this architecture: (a) The MAV autonomously performs a low-altitude search mission under the forest canopy. (b) Each captured frame is processed in real time onboard using an optimized lightweight detection model. (c) Alternatively, the video stream is transmitted over a commercial 5G network to a remote edge server, where inference is performed by a more powerful model.

Field experiments were conducted under canopy conditions using both mannequins and human actors to simulate missing persons. In these trials, both onboard and edge inference modes successfully detected targets in realistic environments, demonstrating the effectiveness of our SAR system and the applicability of ForestPersons to real-world scenarios.

B Benchmark Models

B.1 Implementation details

Table 5: **Hyperparameter settings for training object detections.** Most of the configurations are followed by the default setting provided by MMDetection and detrex.

Methods	Optimizer	Learning rate	Batch size	Weight decay	Epoch
YOLOv3 [30]	SGD	1×10^{-3}	64	5×10^{-4}	273
YOLOX [31]	SGD	1×10^{-2}	64	5×10^{-4}	300
RetinaNet [33]	SGD	5×10^{-3}	16	1×10^{-4}	12
Faster R-CNN [27]	SGD	2×10^{-2}	16	1×10^{-4}	12
SSD [35]	SGD	1.5×10^{-2}	192	4×10^{-5}	120
DETR [37]	AdamW	1×10^{-4}	16	1×10^{-4}	150
DINO [38]	AdamW	1×10^{-4}	16	1×10^{-4}	12

In this section, we describe the hyperparameter settings used to train each object detection model for benchmarking purposes. Table 5 summarizes the configurations for all models. Most hyperparameters follow the default settings provided by the MMDetection [39] and detrex [40] frameworks, except RetinaNet, for which we reduced the learning rate compared to the default setting to prevent training instability observed with higher values.

B.2 Analysis of benchmark models

Given the critical nature of SAR missions, achieving high recall is a primary requirement, necessitating a more deliberate examination of the precision-recall trade-off compared to conventional object detection tasks. To investigate this aspect, we present the precision-recall curve at an IoU threshold of 0.5, as illustrated in Figure 8. The precision-recall curve is constructed by sorting predicted bounding boxes in descending order of confidence scores, with increasing confidence thresholds prioritizing precision over recall, while lower thresholds capture more true positives at the cost of introducing false positives. The resulting shape of the curve characterizes how each model behaves under varying confidence thresholds, offering insight into its sensitivity to recall-focused operating points.

In Figure 8, DINO [38] (pink) consistently maintains high recall even at low confidence thresholds, whereas SSD [35] (purple) exhibits a clear limitation in its recall capacity. Specifically, even when all predicted bounding boxes are treated as true positives, its curve saturates below the recall levels reached by DINO. This indicates a structural limitation in SSD’s detection capability that cannot be overcome by threshold tuning alone. Such findings indicate that, particularly in SAR contexts, the upper bound of recall achievable by a model constitutes an essential metric in itself, complementing traditional aggregate measures such as mAP.

In practice, the confidence threshold is often selected based on the point that maximizes the F1-score, calculated on a validation or a test set. However, in recall-sensitive domains such as SAR, it may be more appropriate to deliberately reduce the threshold to prioritize recall, even at the expense of an increased false positive rate. This strategy aligns with real-world operational considerations, wherein human operators may prefer investigating more candidate detections rather than risking failure to detect actual missing persons. Therefore, we argue that the development and evaluation of object detectors for SAR applications should incorporate not only AP but also (1) the maximum attainable recall and (2) the recall level at which precision begins to decline sharply. These indicators are closely tied to the likelihood of successfully locating and rescuing missing persons, and thus serve as critical performance criteria in SAR applications.

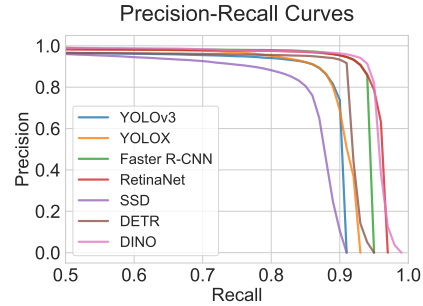


Figure 8: **Precision-Recall curves of baseline object detection models.**

C Case Study: Successes, Failures, and Future Directions

C.1 Limitations of Generalization from Prior Benchmarks

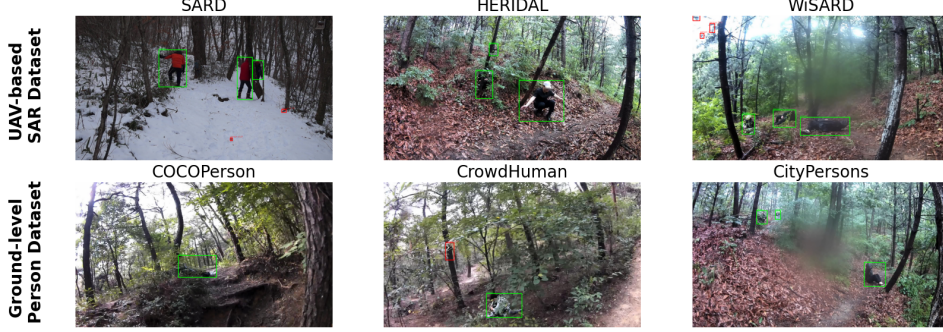


Figure 9: **Failure cases of object detection models trained on prior UAV-based SAR and Ground-level Person datasets.** Green boxes indicate ground-truth bounding boxes, and red boxes represent model predictions. These examples illustrate the limitations of existing datasets in handling under-canopy SAR scenarios.

ForestPersons differs from conventional detection benchmarks in several key aspects, including viewpoint, environmental complexity, and the conditions of human targets. To assess generalizability, we evaluated models trained on existing SAR datasets and ground-level person datasets using the ForestPersons test split. Specifically, we selected SARD [12], HERIDAL [10], and WiSARD [11] as representative UAV-based SAR datasets, and COCOPersons [20], CrowdHuman [21], and CityPersons [22] as representative ground-level person datasets. For all experiments, we used Faster R-CNN [27] as the object detection model.

As illustrated in Figure 9, models trained on these existing datasets exhibit limited generalizability when applied to under-canopy SAR scenarios. This outcome is expected: prior UAV-based SAR datasets primarily contain aerial images, which differ significantly from the ground-level perspectives that are characteristic of under-canopy tasks. While ground-level datasets more closely reflect the viewpoint of MAV flights compared to conventional UAV-based SAR datasets, they still predominantly feature upright and fully visible individuals. Consequently, they fall short in representing challenging cases such as non-standing or heavily occluded persons, which are common in forest search scenarios.

C.2 Evaluation on ForestPersons



Figure 10: **Success cases of object detection models trained on ForestPersons.** Green boxes indicate ground-truth bounding boxes, and red boxes represent model predictions. The models trained with ForestPersons detect the failure case of the models trained with the existing dataset.



Figure 11: **Failure cases of object detection models trained on ForestPersons.** Green boxes indicate ground-truth bounding boxes, and red boxes represent model predictions. Ground-truth instances with high occlusion or small bounding box size tend to be frequently missed by the detection model.

We then evaluated a model trained on the ForestPersons training split to assess the detection performance gains from using data specifically designed to reflect under-canopy SAR conditions. As shown in Figure 10, the Faster R-CNN model trained on ForestPersons successfully detects missing persons that were not captured by models trained on prior UAV-based or ground-level dataset. This provides qualitative evidence that our dataset better suits SAR tasks in under-canopy environments.

We further investigated the factors contributing to prediction failures on the ForestPersons test set, even when using models trained on ForestPersons. Specifically, we analyzed the prediction results of a Faster R-CNN model trained on ForestPersons by visualizing the confusion matrix, as shown in Figure 12. The confusion matrix summarizes all predictions on the test set and reveals that false positives significantly outnumber false negatives.

However, given the critical nature of SAR tasks, where false negatives are significantly more detrimental than false positives, we focused our analysis on ground-truth instances that were classified as false negatives. Figure 11 presents visual examples of these cases. As expected, the model struggled to detect individuals with small bounding boxes or under heavy occlusion by natural obstacles.

GT: Positive	TP: 19794 (74.8%)	FN: 1073 (4.1%)
	FP: 5579 (21.1%)	TN: N/A
GT: Negative	Pred: Positive	Pred: Negative

Figure 12: **Confusion matrix of the object detection model trained with ForestPersons.**

Interestingly, the winter subset yields noticeably fewer false negatives, suggesting that winter images are generally less challenging for the detection model. A plausible explanation is that individuals in winter scenes are more visually salient due to the higher contrast between individuals and the snow-covered background, which facilitates easier detection. This explanation is supported by the experiment in Table 4b, where a model trained exclusively on winter images generalized worse to the test set than a model trained only on summer images. This indicates that winter images may lack sufficient variability to support effective generalization, which is why they are easier for missing person detection, ultimately reducing the likelihood of false negatives. In contrast, summer images, which often contain dense vegetation leading to various occlusions, contribute more to the generalization ability of the model. These qualitative and quantitative findings help us understand the exceptionally low incidence of false negatives in winter images.

C.3 Generalization Failures from Limited Attribute Training

Extending the results shown in Table 4, we further analyzed how restricting training data to specific attributes, such as pose or season, affects the performance of Faster R-CNN. In Figure 13, models trained only on standing poses perform poorly when detecting people in other postures, such as sitting or lying. These models mainly respond to upright shapes, often mistaking vertical objects like tree trunks for people, and failing to detect people who are lying on the ground. This shows that the model has learned to rely too heavily on shape patterns seen during training, and has not been exposed to enough variation in how missing person actually appear during SAR scenarios.



Figure 13: **Detection cases for models trained with data labeled by specific poses.** Green boxes indicate ground-truth bounding boxes, and red boxes represent model predictions. Each row corresponds to a model, and each column corresponds to the ground truth pose in the test image. Models trained on a specific pose often fail to detect individuals in other poses and sometimes identify incorrect regions as humans.

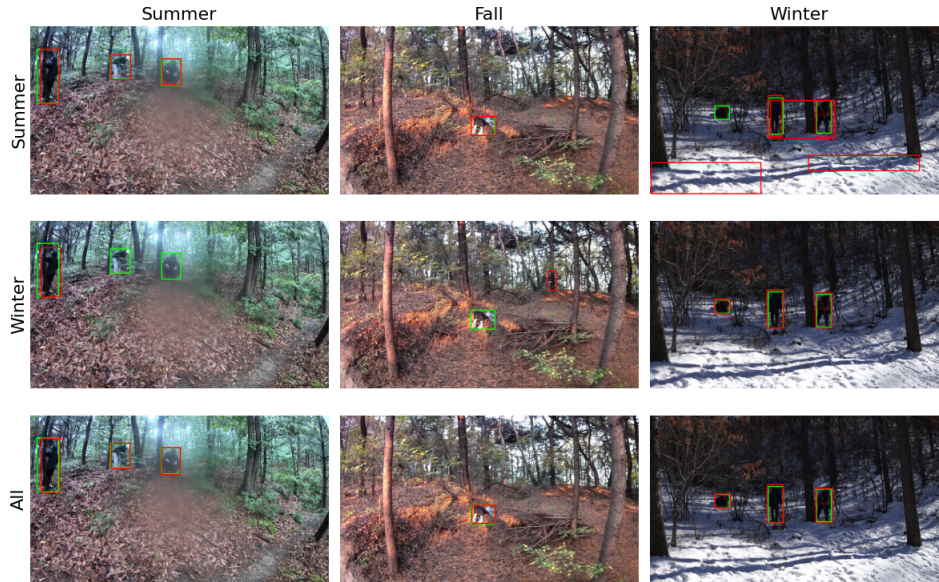


Figure 14: **Detection cases for models trained with seasonal data.** Green boxes indicate ground-truth bounding boxes, and red boxes represent model predictions. Each row shows detection from a model trained on data from a specific season, while each column represents test data from a particular season. Models trained on limited seasonal data show clear seasonal bias when applied to scenes from different seasons, such as failing to detect people or generating inaccurate bounding boxes.

A similar pattern is seen in the seasonal experiments in Figure 14. Models trained only on summer images, which contain more vegetation and frequent occlusion, show slightly better generalization to other seasons. The presence of dense vegetation and natural occlusion in summer scenes appears to help the model learn features that generalize better to different seasonal environments. However, these models still produce many errors in winter scenes, such as false positives caused by mistaking snow-covered terrain for people. In contrast, models trained only on winter images perform significantly worse in other seasons. Winter scenes usually lack vegetation and have fewer occluding elements, which limits the diversity of visual cues the model can learn from. As a result, these models often fail to detect people in summer scenes with dense foliage and complex backgrounds, leading to frequent false negatives. This tendency is reflected in the confusion matrices shown in Figure 15. These findings indicate that the visual properties of each season shape how the model learns and

GT: Positive	TP: 16433 (52.7%)	FN: 4434 (14.2%)	GT: Positive	TP: 12383 (42.5%)	FN: 8484 (29.2%)
	FP: 10312 (33.1%)	TN: N/A		FP: 8237 (28.3%)	TN: N/A
GT: Negative			GT: Negative		
	Pred: Positive	Pred: Negative		Pred: Positive	Pred: Negative

Figure 15: **Confusion matrices of object detection models trained on summer and winter datasets.** (Left) Summer-trained model; (Right) Winter-trained model.

where it tends to fail, and that training on a single season is not sufficient to ensure robustness across seasonal conditions.

Unlike models trained on a single season or pose, the model trained on the complete dataset, which includes a full range of poses and seasonal conditions, performs more reliably, as shown in the last rows of Figure 13 and Figure 14. These results demonstrate the effectiveness of ForestPersons as a benchmark that reflects the diversity and complexity of real-world SAR conditions. By providing extensive variation in human pose, occlusion, and environmental factors, ForestPersons supports the development of more generalizable models and serves as a solid foundation for advancing robust missing person detection in challenging under-canopy search tasks.

C.4 Limitations Exposed and Directions for Future SAR Detection

Our qualitative analysis highlights the utility of ForestPersons in diagnosing the generalization and structural limitations of representative detection models in the context of SAR missions. ForestPersons introduces new challenges by incorporating vegetation-rich environments that frequently cause occlusion, diverse human poses including non-upright postures, and seasonal conditions such as snow that are often absent in prior datasets. These findings show that models trained on narrow visual patterns may seem reliable in simplified test environments but fail to maintain the same level of reliability when applied to real-world conditions. While ForestPersons was carefully designed to cover a wide range of poses, occlusion levels, and seasonal conditions, our analysis suggests that some failure cases may still remain undetected. Dataset diversity is therefore critical for revealing model limitations, but it alone may not be sufficient.

To address this, complementary approaches such as optimizing viewpoint and trajectory design can further reduce the inherent difficulty of the detection task and enhance practical performance in the field. One such approach is viewpoint-aware flight planning, which can support vision models by improving the visibility of missing persons. By explicitly accounting for the MAV’s camera field of view, such planning can help ensure that individuals are captured from favorable angles and distances. In contrast, coarse trajectories that simply follow major roads may expose the model to less informative and more occluded perspectives. Therefore, alongside the use of diverse datasets like ForestPersons, flight strategies that structurally facilitate detection should be explored as a complementary direction, particularly in the context of autonomous SAR missions.

D Data Collection Guidelines

ForestPersons was constructed to reflect realistic search scenarios for missing persons in forested environments. All video sequences were recorded using handheld or tripod-mounted cameras, including GoPro HERO 9 Black, Sony SLT-A57, and See3CAM 24CUG models. The cameras were positioned to simulate the typical flight altitudes and viewing angles of low-altitude MAVs operating

under forest canopy, capturing slightly downward-facing perspectives similar to those used in actual search operations. All recordings were captured at a frame rate of at least 20 FPS, with resolution settings adjusted depending on the camera model used.

D.1 Locations: Forest Environments Relevant to SAR Missions

All data were collected in forested regions where real-world missing person incidents are likely to occur. We selected diverse environments including dense forest interiors, valleys, and forest entrances to reflect typical terrain encountered during SAR missions. These locations span a range of vegetation density and visibility conditions, from heavily occluded forest interiors to forest edge regions with sparse vegetation.

Each environment includes natural sources of visual occlusion such as tree branches, underbrush, uneven terrain, and varying vegetation density. We aimed to incorporate diverse spatial layouts that challenge missing person detection, including not only typical forest trails but also rocky valleys and steep slopes covered with dense foliage. This diversity enables the dataset to capture a broad range of search scenarios encountered in SAR missions.

D.2 Weather and Time of Day

To reflect the environmental diversity encountered in real-world search operations, data were collected under various weather and lighting conditions. All video sequences were captured during daytime or twilight hours before sunset, when there was sufficient natural light. Night time scenes were excluded due to safety concerns during field deployment and the limited effectiveness of RGB-based detection in low-light conditions.

Weather and seasonal conditions included sunny, overcast, and snow-covered winter environments. These variations allowed us to capture diverse visual appearances, including strong shadows under direct sunlight, diffuse lighting on cloudy days, and high reflectance and severe occlusion in snowy terrain. Each sequence is accompanied by metadata describing both the season and weather, enabling evaluations under specific environmental contexts.

D.3 Subject Behavior and Capture Strategy

To simulate realistic SAR scenarios, actors in the ForestPersons performed a wide range of behaviors, including standing, sitting, lying down, and natural transitions between these states. Transitional poses (e.g., moving from a seated to a standing position) were annotated with the nearest posture label, typically sitting or standing. Although this labeling may involve some degree of annotator subjectivity, its impact on the overall data quality is minimal. These behavioral variations reflect the diversity of human configurations encountered in SAR operations.

Camera platforms included handheld rigs and tripods. To emulate the viewpoint of MAVs operating under canopy, operators followed movement paths consistent with low-altitude MAV trajectories. Camera height, angle, and distance were varied within and across sequences to simulate oblique and horizontal viewpoints. This variation allowed us to capture human subjects from perspectives representative of realistic aerial search conditions.

A key aspect of our strategy was the active creation of natural occlusion. Rather than using fixed occlusion setups, camera operators navigated around tree branches, bushes, or through dense vegetation to partially obscure subjects in dynamic and realistic ways. In difficult environments such as snowy or rainy terrain, where operator movement posed safety risks, the camera was fixed and actors moved within the frame to simulate occlusion safely.

E Video Sequence-Level Difficulty Estimation

ForestPersons was collected as a set of video sequences, from which image frames were extracted to construct the final dataset. In this setup, if frames from the same sequence are split across training, validation, and test splits, it can lead to overestimated model performance. This is because detection models may implicitly learn scene-specific backgrounds or appearances during training, and then encounter similar contexts during evaluation, resulting in inflated accuracy that does not reflect true



Figure 16: **Examples across various visibility levels and poses.** Images are grouped by visibility level (rows) and pose (columns), each drawn from distinct scene contexts.

generalization. To avoid such overlap, we split the dataset at the sequence level, ensuring that each video sequence appears in only one of the train, validation, or test splits.

E.1 Necessity of Difficulty-Aware Data Splitting

A naive approach such as randomly assigning sequences to each split, or manually selecting them based on subjective judgment (e.g., "easy-looking" or "challenging" scenes), can lead to distributional bias across splits. For example, one split might inadvertently contain mostly clear and well-lit scenarios, while another might be dominated by occluded or low-visibility scenes. Such imbalance can undermine the fairness and interpretability of model comparisons.

To mitigate this issue, we introduced a model-based method for estimating sequence-level difficulty, providing a principled way to assess and distribute difficulty across the dataset.

E.2 Model-Based Difficulty Estimation

We employed a Faster R-CNN [27] object detector pretrained on the COCO [20] dataset to estimate the detection difficulty of each sequence. For each sequence, we applied the detector to all images and computed the Average Precision (AP). The difficulty score for a sequence s is then defined as:

$$\text{Difficulty}(s) = 1 - \text{AP}_{50}(s) \quad (1)$$

Here, $\text{AP}_{50}(s)$ denotes the performance of the detector model on sequence s , averaged over all annotated frames. Higher AP values indicate that the sequence is easier to detect, while a lower AP

corresponds to more challenging scenes. This formulation provides an objective difficulty measure, independent of annotator intuition or handcrafted heuristics.

E.3 Difficulty-Aware Dataset Splitting

Based on the estimated difficulty scores, we sorted all video sequences in ascending order of AP (i.e., increasing difficulty) and allocated them to train, validation, and test splits to ensure balanced difficulty distribution. For example, sequences were interleaved across splits so that each contained a diverse mixture of easy, medium, and hard samples.

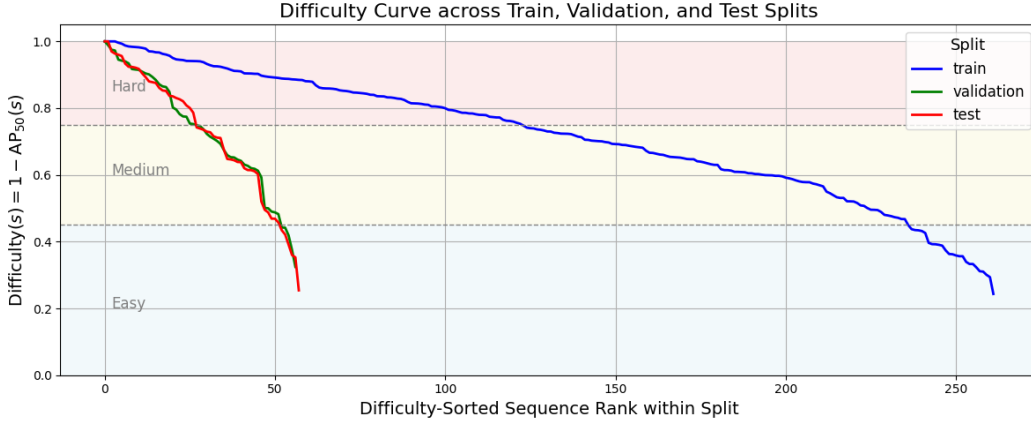


Figure 17: Difficulty Curve across Train, Validation, and Test Splits

As shown in Figure 17, the difficulty curve of ForestPersons illustrates that each sequence spans a range of detection difficulty. Each point corresponds to a video sequence, sorted by its model-based difficulty score $1 - \text{AP}_{50}(s)$. The plot illustrates that the dataset spans a broad range of difficulty levels, ensuring balanced evaluation across splits.

References

- [1] Xu Liu, Guilherme V Nardari, Fernando Cladera, Yuezhan Tao, Alex Zhou, Thomas Donnelly, Chao Qu, Steven W Chen, Roseli AF Romero, Camillo J Taylor, et al. Large-scale autonomous flight with real-time semantic slam under dense forest canopy. *IEEE Robotics and Automation Letters*, 7(2):5512–5519, 2022.
- [2] Abraham Bachrach, Anton de Winter, Ruijie He, Garrett Hemann, Samuel Prentice, and Nicholas Roy. Range-robust autonomous navigation in gps-denied environments. In *2010 IEEE International Conference on Robotics and Automation*, Anchorage, Alaska, US, 2010. IEEE.
- [3] Sebastián Barbas Laina, Simon Boche, Sotiris Papatheodorou, Dimos Tzoumanikas, Simon Schaefer, Hanzhi Chen, and Stefan Leutenegger. Scalable autonomous drone flight in the forest with visual-inertial slam and dense submaps built without lidar. *arXiv preprint arXiv:2403.09596*, 2024.
- [4] Youkyung Hong, Suseong Kim, Youngsun Kwon, Sanghyouk Choi, and Jihun Cha. Safe and efficient exploration path planning for unmanned aerial vehicle in forest environments. *Aerospace*, 11(7):598, 2024.
- [5] Yunfan Ren, Fangcheng Zhu, Guozheng Lu, Yixi Cai, Longji Yin, Fanze Kong, Jiarong Lin, Nan Chen, and Fu Zhang. Safety-assured high-speed navigation for mavs. *Science Robotics*, 10(98):eado6187, 2025.
- [6] Laura Jarín-Lipschitz, Xu Liu, Yuezhan Tao, and Vijay Kumar. Experiments in adaptive replanning for fast autonomous flight in forests. In *2022 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2022.
- [7] Boyu Zhou, Yichen Zhang, Xinyi Chen, and Shaojie Shen. Fuel: Fast uav exploration using incremental frontier structure and hierarchical planning. *IEEE Robotics and Automation Letters*, 6(2):779–786, 2021.
- [8] Tzu-Jui Lin and Karl A. Sto. Autonomous surveying of plantation forests using multi-rotor uavs. *Drones*, 6(9):256, 2022.
- [9] Youngsun Kwon, Suseong Kim, Youkyung Hong, Sanghyouk Choi, and Jihun Cha. Online terrain mapping for exploring dense forests on unmanned aerial vehicles. In *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1676–1680. IEEE, 2024.
- [10] Mirela Kundid Vasić and Vladan Papić. Improving the model for person detection in aerial image sequences using the displacement vector: A search and rescue scenario. *Drones*, 6(1):19, 2022.
- [11] Daniel Broyles, Christopher R Hayner, and Karen Leung. Wisard: A labeled visual and thermal image dataset for wilderness search and rescue. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9467–9474. IEEE, 2022.
- [12] Sasa Sambolek and Marina Ivacic-Kos. Search and rescue image dataset for person detection (sard). *IEEE Dataport*, 2021.
- [13] Xiangqing Zhang, Yan Feng, Nan Wang, Guohua Lu, and Shaohui Mei. Aerial person detection for search and rescue: Survey and benchmarks. *Journal of Remote Sensing*, 5, 2025.
- [14] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7380–7399, 2021.
- [15] Simon Speth, Artur Gonçalves, Bastien Rigault, Satoshi Suzuki, Mondher Bouazizi, Yutaka Matsuo, and Helmut Prendinger. Deep learning with rgb and thermal images onboard a drone for monitoring operations. *Journal of Field Robotics*, 39(6):840–868, 2022.

- [16] Mohammadamin Barekatin, Miquel Martí, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. Okutama-action: An aerial view video dataset for concurrent human action detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 28–35, 2017.
- [17] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018.
- [18] Jingru Zhu, Xiandong Wang, Yi Liu, Qianwei Ji, Zhao Zhao, and Shengke Wang. Uavtinydet: Tiny object detection in uav scenes. In *2022 7th International conference on image, vision and computing (ICIVC)*, pages 195–200. IEEE, 2022.
- [19] Yingying Liu, Fengqin Yao, Laihui Ding, Zhiwei Xu, Xiaogang Yang, and Shengke Wang. An image segmentation method based on transformer and multi-scale feature fusion for uav marine environment monitoring. In *2023 8th International Conference on Image, Vision and Computing (ICIVC)*, pages 328–336. IEEE, 2023.
- [20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.
- [21] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.
- [22] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3213–3221, 2017.
- [23] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3354–3361. IEEE, 2012.
- [24] Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Sheno, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):6748–6765, 2021.
- [25] Justin Brooks. Coco annotator. <https://github.com/jsbrooks/coco-annotator/>, 2019.
- [26] Josep López. YOLOv8-Face: A yolov8-based face detection implementation. <https://github.com/Yusepp/YOLOv8-Face>, 2024.
- [27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
- [28] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019.
- [29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.
- [30] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
- [31] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.
- [32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

- [33] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 2980–2988, 2017.
- [34] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [35] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *Proceedings of the European conference on computer vision (ECCV)*, page 21–37, 2016.
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [37] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [38] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.
- [39] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.
- [40] Tianhe Ren, Shilong Liu, Feng Li, Hao Zhang, Ailing Zeng, Jie Yang, Xingyu Liao, Ding Jia, Hongyang Li, He Cao, Jianan Wang, Zhaoyang Zeng, Xianbiao Qi, Yuhui Yuan, Jianwei Yang, and Lei Zhang. detrex: Benchmarking detection transformers, 2023.