Elevating point-based object detection in UAVs: A deep learning method with altitude fusion

Michał Wiliński[0009−0004−4818−8417], **Bartosz Ptak**[0000−0003−1601−6560], **Marek Kraft**[0000−0001−6483−2357]

Poznan University of Technology
Faculty of Control, Robotics and Electrical Engineering
Piotrowo 3A, 60-965 Poznań, Poland
marek.kraft@put.poznan.pl

Abstract. *Recent advancements in computer vision and deep learning have revolutionised remote sensing. An important challenge lies in detecting small objects like individuals in crowds from low-altitude drone-captured images, which are problematic due to scale variations. While existing research addresses the scale challenges, not enough focus has been put on the exploitation of altitude data from UAV sensors. This paper proposes three deep learning-based methods to fuse altitude and Ground Sampling Distance (GSD) information, which enhance input images with additional features and showcase significant improvements in point-oriented object detection on our custom dataset.*
Keywords: *remote sensing, deep learning, GSD and altitude fusion, data fusion*

1. Introduction

In recent years, convolutional neural networks (CNNs) and deep learning have significantly surpassed traditional computer vision algorithms in terms of performance. This progress is also notable in remote sensing tasks involving images captured by Unmanned Aerial Vehicles (UAVs). Researchers are increasingly focusing on the development of these methods due to their applicability across various domains, such as road safety [1] and Smart Cities initiatives [2]. One key area of focus involves the detection of tiny objects, such as individual persons

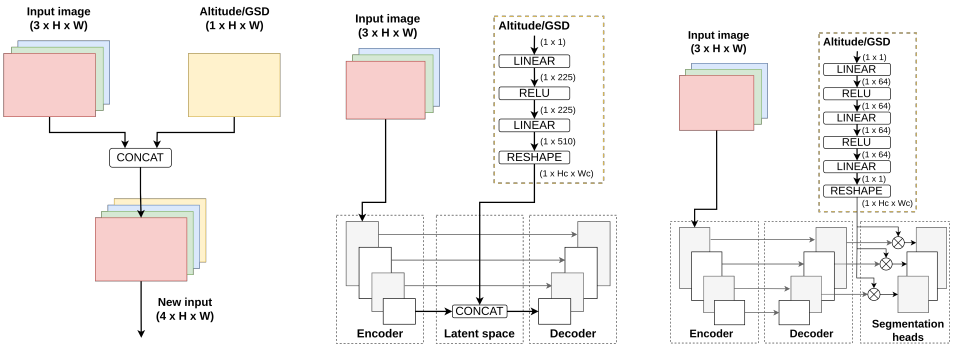
within a crowd, using low-altitude aerial imagery captured by drones. Unlike typical closed-circuit television (CCTV) applications, algorithms for drone imagery analysis must effectively address challenges like scale variations, perspective distortions, and movement-induced changes. Previous research has proposed various methods to address the performance drop due to scale variability of objects in remote-sensing images. However, none of these methods have leveraged direct altitude information from UAV sensors to enhance the task of detecting tiny objects, which may come as a surprise, since this information is usually directly and easily available.

In this paper, we propose and compare three methods dedicated to fusing altitude information, thereby enhancing input images with additional features. For each method, we evaluate the effectiveness of both raw altitude information and Ground Sampling Distance (GSD) to determine the most efficient fusion approach. Our experiments on the private dataset demonstrate a significant improvement compared to the baseline model in the task of detecting point-oriented objects.

2. Related Work

The integration of numerical information with image data, often referred to as data fusion, has been a subject of interest in various fields. One notable article that explores this intersection is [3]. This study explores different methodologies for integrating electronic health records (EHR), consisting of numerical and categorical features, with medical imaging. The authors identify and discuss three primary fusion strategies: early fusion, performed before inputting data into the model; joint fusion, where intermediate representations from different modalities are combined; and late fusion, which involves aggregating predictions from modality-specific models. Another work [4] proposes attentional feature fusion as a scheme for integrating features in modern network architectures, addressing issues related to inconsistent semantics and scales.

In the field of remote sensing, numerous studies have delved into data fusion. In [5], the authors analysed the data fusion capabilities for multisource data, concentrating solely on fusing images from different sensors. Another instance of data fusion is demonstrated in [6], where a two-step approach is employed. Initially, a deep learning model is trained for GSD estimation using images. Subsequently, the feature vector of the model is combined in the latent space to enhance the object detection task. Unexpectedly, none of these methods explored the direct impact of fusing altitude and GSD information available directly from the sensors.



(a) Method A - fuse raw information as an additional input channel.

(b) Method B - fuse as learnable features in latent space.

(c) Method C - fuse as learnable features at decoder stages level.

Figure 1: Overview of altitude information fusion methods. ALT - altitude information; GSD - Ground Sampling Distance.

3. Methods

Following state-of-the-art methods for point-oriented object detection, we employ the UNet [7] architecture with the EfficientNet-b2 [8] backbone, including weights pretrained on the ImageNet dataset. This model processes input images to produce an output mask where non-zero values represent estimated detection coordinates. Furthermore, to enhance the training process, we implement Deep Supervision, a widely recognised technique in remote sensing renowned for its ability to capture fine-grained details. The model takes images and resizes them to the resolution of 960×544 and generates masks in the same resolution, which are interpolated to the original resolution in the final step. The AdamW optimiser with the changing learning rate was applied in the training process, wrapping in the CosineAnnealingLR scheduler with an initial learning rate of $3e^{-4}$. We also perform augmentations (Flip, Rotate, Noise, Colour shift, Random Gamma) to improve the model's generalisation capabilities using varied images.

This architecture and training procedure are used as a baseline for our study. The methods outlined below follow the same training procedure, differing only in architecture and fusion methodology. Fig. 1 provides a general overview of the proposed methods.

Method A. In this approach, an additional channel is concatenated to the input

image, wherein the values represent normalised altitude or GSD. This approach stands out as the most efficient and straightforward means of integrating numerical data with input images. A visual representation of this process is depicted in Fig. 1a.

Method B. An alternative approach involves manipulation within the model's latent space (Fig. 1b). In this method, we suggest the incorporation of a learnable altitude or GSD embedding with a single latent channel size. Subsequently, this embedding undergoes resizing to align with the dimensions of the corresponding feature maps at that level and is appended as an additional feature map. This adaptation facilitates the development of a trainable altitude/GSD representation.

Method C. Finally, we focused on the intersection of the decoder and segmentation heads. As shown in Fig. 1c, we augmented the output of each decoder stage by performing element-wise product between output feature maps and learnable embedding of normalised GSD or altitude values. Subsequently, these augmented feature maps serve as inputs to the segmentation heads.

4. Evaluation

4.1. Dataset

We assess the method using an in-house dataset comprising 10000 images captured by the DJI Mini 2 drone at various altitudes ranging from 26.0 to 101.0 meters, with an average altitude of 60.3 meters. The dataset provides a list of coordinates delineating the human head for each image, along with corresponding altitude information indicating the capture height and is divided into train (6893), validation (1391), and test (1716) splits. Thanks to known altitude and camera parameters, we can calculate the GSD. We illustrate sample images featuring labelled heads in Fig. 2.

4.2. Results

In assessing the compared methods, we employ point-oriented detection metrics including Precision, Recall, and F1-Score, which are computed for test subsets. It is crucial to highlight that accurate detection is defined by an Euclidean distance between points that is equal to or less than 5 pixels in original size.

The experiment results are placed in Tab. 1. In general, the Ground Sample Distance (GSD) consistently outperformed altitude in terms of overall results. De-

Figure 2: Two example images with drawn head labels (red circle) on extremes of altitudes (left: 26.0m; right: 101.0m).

spite altitude exhibiting a marginally higher precision of 0.731, the GSD demonstrated better recall with a score of 0.737. Method A and B emerged as the frontrunners in this study, showcasing the best overall F1-score of 0.712. Furthermore, Method B's performance in achieving the highest F1-score for altitude reached a value of 0.711. All methods demonstrated a notable increase in performance compared to the baseline, indicating that the fusion of altitude or GSD data contributes to enhancing the model's performance.

Table 1: Methods comparison considering raw altitude and GSD fusion.

	Altitude			**GSD**		
Method	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Baseline	0.697	0.725	0.701	0.697	0.725	0.701
Method A	0.721	0.701	0.705	0.698	0.737	**0.712**
Method B	0.731	0.703	0.711	0.700	0.734	**0.712**
Method C	0.685	0.737	0.705	0.712	0.718	0.710

5. Conclusions

In this paper, we present that integrating altitude data with input imagery enhances performance metrics for detecting point-oriented objects in low-altitude aerial images. Moreover, our experiments reveal that combining Ground Sampling Distance (GSD) achieves better results compared to using altitude alone. We also show that the fusion implemented at the encoder is more effective rather than at the decoder level.

Data availability. We declare the dataset is not published within the article.

References

[1] Outay, F., Mengash, H. A., and Adnan, M. Applications of unmanned aerial vehicle (UAV) in road safety, traffic and highway infrastructure management: Recent advances and challenges. *Transportation research part A: policy and practice*, 141:116–129, 2020.

[2] Kharchenko, V., Kliushnikov, I., Rucinski, A., Fesenko, H., and Illiashenko, O. UAV fleet as a dependable service for smart cities: Model-based assessment and application. *Smart Cities*, 5(3):1151–1178, 2022.

[3] Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., and Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *Digital Medicine*, 3, 2020.

[4] Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., and Barnard, K. Attentional feature fusion. In *2021 IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 3559–3568. 2021.

[5] Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., Bovolo, F., Chi, M., Anders, K., Gloaguen, R., et al. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 7(1):6–39, 2019.

[6] Yang, Y., Wang, C., Cai, Z., Song, P., Huang, G., Cheng, M., and Zang, Y. GSDDet: Ground sample distance guided object detection for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[7] Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Intern. conf. on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[8] Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conf. on machine learning*, pages 6105–6114. PMLR, 2019.