Towards Scalable and Fast UAV Deployment

Tim Felix Lakemann[a] and Martin Saska[b]
Department of Cybernetics, Czech Technical University, Karlovo namesti 13, Prague, Czech Republic

Keywords: Object Detection, Segmentation and Categorization, Omnidirectional Vision, Multi-Robot Systems, Computer Vision for Automation.

Abstract: This work presents a scalable and fast method for deploying Uncrewed Aerial Vehicle (UAV) swarms. Decentralized large-scale aerial swarms rely on onboard sensing to achieve reliable relative localization in real-world conditions. In heterogeneous research and industrial platforms, the adaptability of the individual UAVs enables rapid deployment in diverse mission scenarios. However, frequent platform reconfiguration often requires time-consuming sensor calibration and validation, which introduces significant delays and operational overhead. To overcome this, we propose a method that enables rapid deployment and calibration of vision-based UAV swarms in real-world environments.

1 INTRODUCTION

Collaborative multi–UAV systems improve robustness and operational efficiency across a wide range of applications, from search and rescue to environmental monitoring (Bartolomei et al., 2023). Accurate relative localization of team members is essential for safe navigation, collision avoidance, and coordinated task execution (Chung et al., 2018; Chen et al., 2022). Global Navigation Satellite System (GNSS) alone is often insufficient for these tasks due to its limited precision, unavailability in GNSS–denied environments, and vulnerability to interference (Xu et al., 2020; Zhou et al., 2022). Although alternatives such as Real Time Kinematics (RTK)-GNSS and motion capture systems provide high accuracy, they require external infrastructure or connectivity, making them unsuitable for many real-world scenarios (Chung et al., 2018).

In swarm applications, vision-based relative localization systems onboard offer scalable, cost-effective, and decentralized solutions to detect and track other UAVs, as shown in recent studies (Li et al., 2023; Zhao et al., 2025).

From our experience deploying large-scale swarms with diverse robot configurations and sensors, we identified sensor calibration as a primary bottleneck for fast real-world deployment. In both industrial and research settings–such as object detection, terrain analysis, or communication–aware navigation-sensor changes are common and often necessitate recalibration. Although recent vision-based methods based on deep learning (Schilling et al., 2021; Xu et al., 2020; Oh et al., 2023) have shown promise, Convolutional Neural Networks (CNNs) are typically computationally intensive, require large annotated datasets, tend to overfit to specific platforms, and struggle to generalize to unseen conditions like sun reflections (Funahashi et al., 2021). Additionally, visible parts of UAV—-such as the rotor arms, landing gear, or camera mount–can introduce artifacts, degrade algorithm performance, or require manual pre-processing. Such artifacts can lead to false detections or misclassifications, causing onboard systems to make incorrect decisions. In the worst case, this can result in collisions with obstacles or other agents in a collaborating swarm. One possible mitigation strategy is to mount the camera so that no part of the UAV frame appears in Field of View (FOV); however, this is often impractical for small aerial vehicles or when omnidirectional vision

Figure 1: Overview of collaborating UAVs using our prior auto-generated masks for save collaboration.

[a] https://orcid.org/0009-0000-4863-3235
[b] https://orcid.org/0000-0001-7106-3816

Lakemann, T. F. and Saska, M.
Towards Scalable and Fast UAV Deployment.
DOI: 10.5220/0013713500003982

In *Proceedings of the 22nd International Conference on Informatics in Control, Automation and Robotics (ICINCO 2025) - Volume 1*, pages 243-250
ISBN: 978-989-758-770-2; ISSN: 2184-2809

is required. In addition, off-center mounting can shift the center of mass, compromising flight stability, particularly critical for agile or tightly constrained UAV platforms.

An effective solution to mitigating reflections and other artifacts caused by the UAV frame in camera images is to generate a mask that excludes these regions. However, manually annotating such frame regions introduces a significant bottleneck, hindering scalability and delaying the deployment of perception pipelines across different UAV platforms. The problem of automatically identifying and masking the visible UAV frame has received limited attention in the literature and, to the best of the authors' knowledge, is often neglected entirely, risking the misclassification of structural elements of the observing UAV–or handled manually.

In this work, we propose a novel approach for automatic detection and masking of the UAV frame in on-board imagery. This method serves as an enabling technology for the rapid preparation and deployment of large-scale aerial swarms (Fig. 1). It is lightweight, does not require prior training, and is adaptable to various camera models and UAV configurations. The approach supports user interaction and validation, producing high-quality masks that effectively exclude visible UAV structures, thereby facilitating faster deployment while improving the safety and reliability of swarm operations.

The proposed approach is open-source and available at https://github.com/ctu-mrs/uvdar_core/blob/master/scripts/extract_mask.py.

2 STATE OF THE ART

The automatic mask generation to extract the frame of UAV is a problem not present in the current literature. However, object detection algorithms have been employed to identify and mask unwanted elements in UAV imagery. For example, in (Pargieła, 2022), the authors used YOLOv3 to detect and mask vehicles in images acquired from UAV, thereby enhancing the quality of digital elevation models and orthophotos.

CNNs have been widely adopted for semantic segmentation. These models assign class labels to each pixel, facilitating a detailed understanding of the scene. For example, in (Soltani et al., 2024), the authors demonstrated the efficacy of segmentation based on CNN in orthoimagery UAV for the classification of plant species. They use a two–step approach. First they trained a CNN-based image classification model using simple labels and applied it in a moving-window approach over UAV orthoimagery to create segmentation masks. In the second phase, these segmentation masks were used to train state-of-the-art CNN-based image segmentation models with an encoder–decoder structure.

However, training such models requires labeled datasets, which are often scarce in UAV applications due to the labour-intensive nature of manual annotation. To mitigate this, researchers have explored the use of synthetic data. In (Hinniger and Rüter, 2023), the authors generated synthetic training data using game engines to simulate UAV perspectives, thereby augmenting real datasets and improving model performance. The recent work of (Maxey et al., 2024) introduces a simulation platform for UAV perception tasks using Neural Radiance Fields (NeRF) (Mildenhall et al., 2021). While the focus is on generating synthetic datasets for tasks like obstacle avoidance and navigation, a key contribution is the explicit 3D modeling of the UAV structure and camera perspectives, which inherently involves managing occlusions and self-visibility. Mask-guided techniques have also been explored in generative tasks. For instance,in (Zhou et al., 2024), the authors used semantic masks to control UAV-based scene synthesis, showcasing the broader applicability of mask-based conditioning in UAV imagery.

Recent advancements have seen the integration of transformer architectures in UAV image segmentation. Models like the Aerial Referring Transformer (AeroReformer) (Li and Zhao, 2025) and Pseudo Multi-Perspective Transformer (PPTFormer) (Ji et al., 2024) have been proposed to address the unique challenges posed by UAV imagery, such as varying perspectives and scales. AeroReformer leverages vision-language cross-attention mechanisms to enhance segmentation accuracy, while PPTFormer introduces pseudo-multiperspective learning to simulate diverse UAV viewpoints.

2.1 Contributions

While existing works primarily focus on the detection and segmentation of external objects, we address the less-explored task of automatically detecting and masking the UAV frame within onboard imagery. Accurate extraction of the UAV frame enables more reliable downstream tasks such as object detection, tracking, and relative localization, by preventing false detections on the UAV's own structure. The main contributions of our work are as follows:

1. We propose a novel method for the automatic detection and masking of the UAV frame in onboard UAV imagery.
2. Our approach incorporates camera-specific

heuristics and leverages spatial relationships via a k-d tree structure, resulting in a lightweight and adaptable solution that generalizes across different UAV platforms and camera configurations.

3. The proposed user-interactive mask generation pipeline does not require prior training or labeled datasets, enabling rapid deployment and customization without the need for supervised learning.
4. Automation of this process facilitates the efficient preparation and deployment of large-scale UAV swarms independent of the UAV platform.

3 METHOD

The method presented in this paper enables a user-interactive mask generation pipeline that extracts the UAV frame from camera images, improving both the faster deployment of UAV swarms and safety. The goal is to automatically detect the frame of the UAV and mask out the frame of the UAV without any user editing. Our method is not limited to a specific system, but applies to any system that requires mask generation of shining parts, usually part of the UAV carrying the camera.

3.1 Image Acquisition

To achieve consistent results, the exposure time and gain settings are fixed, but can be easily adjusted. A dark floor as well as a bright UAV frame is beneficial but is not explicitly required. Further, a dominant light source should be placed above the UAV to ensure visibility of the frame of the UAV.

3.2 Automated Mask Generation

Automated mask generation is the core of the proposed method. In case the image is not a grayscale image, the image would need to be converted to grayscale. Therefore, the grayscale image is denoted by

$$I : \{0,\dots,H-1\} \times \{0,\dots,W-1\} \to \{0,\dots,255\} \tag{1}$$

with H denoting the height of the image and W the width of the image, where $I(y,x)$ denotes the intensity of the pixels with $y \in \{0,\dots,H-1\}$ and $x \in \{0,\dots,W-1\}$. Each candidate pixel is checked using camera-specific constraints to suppress irrelevant regions and reduce false-positive detections. These heuristics depend on the physical location of the camera (e.g. left, right, or back on the UAV) and aim to eliminate known areas with ambient reflections or artifacts.

To reduce false positives due to ambient reflections, we define a camera-specific geometric heuristic:

$$\mathcal{H}_{\text{cam}} \subseteq \{0,\dots,H-1\} \times \{0,\dots,W-1\} \tag{2}$$

which encodes regions of the image known to produce spurious reflections.

We define the filtered domain as:

$$\mathcal{D}' = (0,\dots,H-1 \times 0,\dots,W-1) \setminus \mathcal{H}\text{cam}, \tag{3}$$

and restrict the image I to this domain:

$$I' = I|\mathcal{D}'. \tag{4}$$

Our mask generation algorithm is therefore only applied to I'. In I' pixel values exceeding a threshold σ are considered as potential regions of the UAV frame and stored in the set, denoted by:

$$\mathcal{P} = \left\{(y,x) \in \mathcal{D}' \mid I'(y,x) > \sigma\right\}. \tag{5}$$

Further,

$$\eta < |\mathcal{P}|, \tag{6}$$

must be satisfied, with η denoting the minimum detected points in the camera image. By discarding these dark images, potentially fine masks are not created.

To improve computational efficiency, the points in $\mathcal{P}$ are stored in a k-d tree $\mathcal{T}$ with leaf size l. The tree $\mathcal{T}$ is then used to find the k nearest neighbors of each point in $\mathcal{P}$. For each point $\mathbf{p} \in \mathcal{P}$, the Euclidean distance to each of its k nearest neighbors $\mathbf{q} \in \mathcal{N}_k(\mathbf{p})$ is computed as

$$d(\mathbf{p},\mathbf{q}) = \|\mathbf{p}-\mathbf{q}\|_2, \tag{7}$$

where $\mathcal{N}_k(\mathbf{p})$ denotes the set of k nearest neighbors of $\mathbf{p}$. For each $\mathbf{q} \in \mathcal{N}_k(\mathbf{p})$, if

$$d(\mathbf{p},\mathbf{q}) < \tau, \tag{8}$$

then, the pixel $\mathbf{q}$ and the pixel $\mathbf{p}$ are considered to belong to the same mask in the image, and $\mathbf{q}$ is added to the mask set $\mathcal{M}_\mathbf{p}$ associated with $\mathbf{p}$. The collection of all such mask sets is denoted by

$$\mathcal{M} = \{\mathcal{M}_\mathbf{p} \mid \mathbf{p} \in \mathcal{P}\}. \tag{9}$$

The sets in $\mathcal{M}$ are filled with black color.

To approximate spatial groupings of bright points, each set $\mathcal{M}_\mathbf{p}$ is used to construct a polygon. The points in $\mathcal{M}_\mathbf{p}$ are not explicitly ordered, which may lead to self-intersecting polygons. These polygons are rasterized and used to produce a binary mask in which each

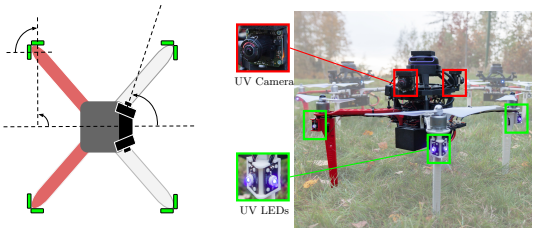


Figure 2: The UVDAR system attached to a UAV and used in the experiments (Licea et al., 2023).

region is filled with black. Overlaps between different sets $\mathcal{M}_\mathbf{p}$ are not explicitly handled; hence, intersecting areas may be filled multiple times, resulting in a binary mask. Once all regions have been filled, the contours are extracted from the previously generated binary mask. These contours are then explicitly drawn on the same mask image with a specified line thickness (κ), reinforcing the region boundaries in the binary mask. The union of all rasterized and contoured polygons is denoted as $\text{poly}(\mathcal{M}_\mathbf{p})$, and the final binary mask image Imasked is defined as:

$$I_{\text{masked}}(\mathbf{x}) = \begin{cases} \text{black}, & \text{if } \mathbf{x} \in \bigcup_{\mathbf{p}\in\mathcal{P}} \text{poly}(\mathcal{M}_\mathbf{p}) \\ \text{white}, & \text{otherwise} \end{cases}. \quad (10)$$

3.3 Interactive Mask Generation

Since validating the masks after creation is essential, a lightweight user interface is provided to support real-time inspection. The user monitors the live stream from the camera and, when the mask creation is activated, the generated binary mask I_{masked}, the original image I, and their overlay I_{overlay} are displayed. The overlay image I_{overlay} is calculated as a visual combination of the original image and the binary mask:

$$I_{\text{overlay}} = \texttt{overlay}(I, I_{\text{masked}}), \quad (11)$$

where `overlay` denotes a pixel-wise operation that highlights masked regions on top of the original image for visualization. This allows the user to immediately assess whether the mask is satisfactory or if it should be discarded and regenerated.

4 EVALUATION

To evaluate our proposed mask generation system, we used the UltraViolet Direction And Ranging (UVDAR) system, a mutual relative localization framework for UAV swarms that operates in the Ultra Violet (UV) spectrum (Fig. 2) (Walter et al., 2019; Horyna et al., 2024; Walter et al., 2018). In

(a) (b)

Figure 3: (a) Side view of the UAV frame, and (b) top view of the UAV frame used in the experiments Two cameras with approximately a FOV of 180 degrees.

Table 1: Parameter settings used during the experiments.

Parameter	Value	Description
σ	120	Binarization threshold
η	20	Min. number of detected points
l	10	Leaf size of $\mathcal{T}$
κ	50	Border Thickness
k	20	nearest neighbors
τ	50	max. distance between neighbors

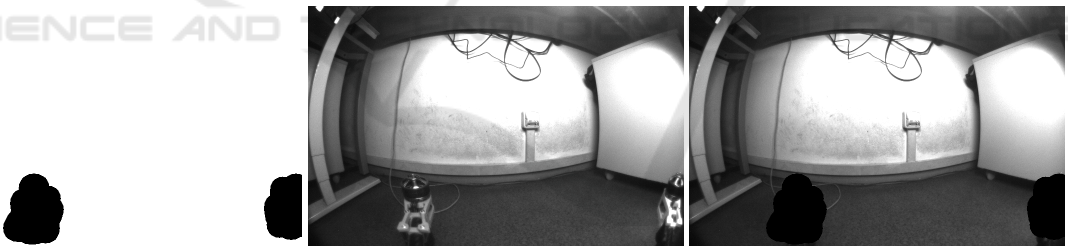
the UVDAR system, cameras are equipped with UV bandpass filters (Walter et al., 2018), which significantly attenuate visible light and allow only UV light to pass through. The swarm members are equipped with UV-Light Emitting Diodes (LEDs) that blink in predefined sequences, enabling the unique identification of each UAV within the swarm. However, the UV light emitted by these LEDs can reflect off the structure of the observing UAV, potentially causing false detections in the camera image. As a result, accurately masking the own structure of the UAV becomes a critical requirement for reliable operation of the UVDAR system. In typical indoor environments such as offices–where ambient UV illumination is minimal–the captured images appear predominantly dark. Therefore, to ensure proper scene visibility, the user was holding a strong UV light source above the observing UAV, as described in Section 3.1, thereby illuminating the environment for effective mask generation (see Figs. 4a and 4b). Grayscale cameras operating under adequate ambient lighting conditions do not require such additional UV illumination.

The evaluation was conducted on two UAV frames using different exposure settings, as detailed in Tab. 2. The high rotational speed of the propellers prevents them from generating noticeable reflections in the camera image. Therefore, they were removed during our mask-generation process. Fig. 3 shows a UAV

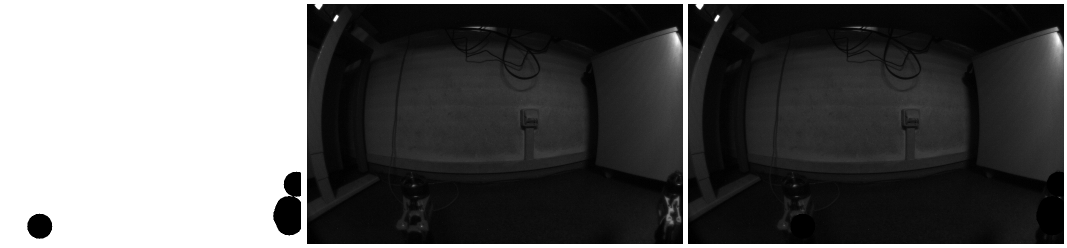
(a) Exp. 1: Sample from UAV1, right camera, with exposure setting: 5000 μs.

(b) Exp. 2: Sample from UAV1, left camera, with exposure setting: 5000 μs.

(c) Exp. 3: Sample from UAV2 right camera, with exposure setting: 25000 μs.



(d) Exp. 4: Sample from UAV2, right camera, with exposure setting: 5000 μs.



(e) Exp. 5: Sample from UAV2, right camera, with exposure setting: 5000 μs.

Figure 4: Qualitative analysis of the generated segmentation masks from different experimental trials. Left to right: generated mask (I_{masked}), input image (I), and overlay image (I_{overlay}). Figures 4a–4d demonstrate successful mask extraction, while Fig.4e shows a failure case under identical conditions to Fig.4d, caused by the removal of the external light source.

Table 2: Average Area, Perimeter, and Compactness with standard deviation for each camera (left or right), exposure time (μs), and UAV for each mask in the image (left and right).

				Area [px^2]		Perimeter [px]		Compactness	
				left	right	left	right	left	right
Exp.	UAV	Cam	Exposure [μs]	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$
1	1	right	5000	8199 ± 619	11477 ± 1042	400 ± 29	471 ± 30	19.6 ± 2.3	19.3 ± 0.9
2	1	left	5000	5901 ± 438	8931 ± 568	317 ± 16	401 ± 13	17.1 ± 0.8	18.0 ± 1.2
3	2	right	25000	11268 ± 2808	7013 ± 1019	458 ± 74	342 ± 26	18.9 ± 2.1	16.8 ± 0.5
4	2	right	2500	9147 ± 3172	5479 ± 1567	392 ± 88	293 ± 49	17.4 ± 1.6	16.0 ± 0.6
5	2	right	5000	7486 ± 2206	6992 ± 1943	393 ± 69	383 ± 88	21.8 ± 4.4	21.4 ± 5.0

frame used for evaluation, equipped with two cameras oriented 140 degrees apart, each with a FOV of 180 degrees. In total, we tested our approach on three BlueFOX-USB cameras mounted on the two UAVs. The evaluation is divided into two parts: quantitative and qualitative analysis.

The quantitative analysis focuses on the consistency and variability of the segmentation masks generated across different cameras and exposure settings. This includes analyzing geometric properties such as the area, perimeter, and compactness of the segmented regions. Qualitative analysis, on the other hand, evaluates the visual quality of the masks and their ability to accurately exclude the UAV frame.

In the context of the UVDAR system, a horizontal cut is applied to each frame to separate the upper and lower image regions. This is necessary because the upper part of the image may contain parts of the light source illuminating the image, which should be excluded from the analysis. We define the top and bottom regions of the image as follows:

$$I_{\text{top}}(y,x) = I(y,x), \quad \text{for } y \in \left[0, \left\lfloor \frac{3}{4}H \right\rfloor - 1\right] \cap \mathcal{D}', \tag{12}$$

$$I_{\text{bottom}}(y,x) = I(y,x), \quad \text{for } y \in \left[\left\lfloor \frac{3}{4}H \right\rfloor, H-1\right] \cap \mathcal{D}'. \tag{13}$$

The segmentation process is designed to run offboard, on a computer with graphical output capabilities. This setup is usually not feasible or cumbersome on lightweight onboard computers such as the NVIDIA Jetson or Intel NUC platforms commonly used on UAVs.

4.1 Mask Shape Analysis via Compactness Metrics

To quantitatively assess the consistency and quality of the generated segmentation masks, we analyzed their geometric properties under varying exposure settings, cameras, and UAVs. In total, we evaluated 70 masks, each generated under different conditions.

We applied connected component analysis to each binary mask after filtering out the UAV frame, identifying valid contiguous regions. For each component, we computed three key shape descriptors: area, perimeter, and compactness, where compactness is defined as:

$$\text{compactness} = \frac{\text{perimeter}^2}{\text{area}}. \tag{14}$$

This metric captures the shape regularity of a component, with lower values generally indicating more compact, circular shapes, and higher values reflecting more elongated or irregular regions.

Masks were categorized by UAV, camera orientation (left or right), and exposure time. Since each UAV included masked regions for both the left and right arms, we evaluated these regions separately. For each category, we computed the mean and standard deviation of the area, perimeter, and compactness in all valid components, as shown in Table 2.

We observed that masks generated under low exposure settings exhibited higher standard deviations, particularly in Experiment 4 (UAV 2, right camera, 2500 μs exposure time). This increased variability indicates a greater sensitivity to lighting conditions, where reduced exposure results in noisier and less consistent segmentations. In contrast, masks captured with exposure times of 5000 μs and 25000 μs showed lower variation and more stable compactness values, suggesting improved reliability in mask generation. Among these, an exposure time of 5000 μs provides a favorable balance between noise suppression and segmentation quality.

4.2 Qualitative Analysis

The qualitative analysis focused on evaluating the visual quality of the generated segmentation masks.

Figure 4 presents five examples of mask images (I_{masked}) under varying exposure settings, environmental conditions (background), and camera configurations, along with the corresponding input image (I) and overlay image I_{overlay}. Our approach successfully extracted the UAV frame across all scenarios tested. As illustrated in Fig. 4e, mask generation failed in the absence of the artificial light source, highlighting the sensitivity of the method to scene illumination. However, as shown in Fig. 4d, the algorithm successfully extracted the UAV frame using the same camera and setup, the only difference being the positioning of the light source. This underscores the importance of maintaining consistent and adequate lighting conditions during mask generation to ensure reliable results.

4.3 Limitations

The generated segmentation masks are generally of good quality, with relatively few false positives and false negatives, although some noise or artifacts may still be present. To further improve robustness, the algorithm should be evaluated across a wider range of frame types and color variations. In low-light conditions, detection of the UAV frame can become more challenging, which can occasionally lead to missed detections.

4.4 Future Work

The proposed algorithm successfully generated masks for the UAV frame in all evaluated scenarios. For future work, we plan to explore the use of an unsupervised Vision Transformer (ViT) to learn the structural characteristics of different UAV frames and enable automatic mask generation. Specifically, our goal is to employ a self-supervised learning framework based on a student-teacher architecture, which can learn robust representations of the frame structure without the need for labeled data. This would further improve the adaptability and scalability of the masking process across varying UAV configurations.

5 CONCLUSIONS

This work introduced a novel method for automatically detecting and masking the frame of a UAV in the images of the camera onboard. By excluding the structure of the UAV in the onboard camera, the method reduces the risk of misclassification, preventing parts of the UAV from being interpreted as obstacles or other agents in multi-UAV systems. This automation significantly enhances the scalability and efficiency of swarm deployment, addressing a task that is otherwise labour-intensive and does not scale when done manually. Leveraging camera-specific geometric heuristics and a k-d tree structure, the proposed algorithm achieves accurate and efficient frame detection across varying UAV designs and camera configurations. Quantitative and qualitative results across multiple platforms confirm the adaptability and robustness of the method in various operating conditions. In general, the proposed approach improves the safety of swarm navigation and onboard vision systems, while also streamlining the preparation and deployment of multi-UAV systems.

ACKNOWLEDGEMENTS

This work was funded by CTU grant no SGS23/177/OHK3/3T/13, by the Czech Science Foundation (GAČR) under research project no. 23-07517S and by the European Union under the project Robotics and advanced industrial production (reg. no. CZ.02.01.01/00/22_008/0004590).

REFERENCES

Bartolomei, L., Teixeira, L., and Chli, M. (2023). Fast multi-uav decentralized exploration of forests. *IEEE Robotics and Automation Letters*, 8(9):5576–5583.

Chen, S., Yin, D., and Niu, Y. (2022). A survey of robot swarms' relative localization method. *Sensors*, 22(12).

Chung, S.-J., Paranjape, A. A., Dames, P., Shen, S., and Kumar, V. (2018). A survey on aerial swarm robotics. *IEEE Transactions on Robotics*, 34(4):837–855.

Funahashi, I., Yamashita, N., Yoshida, T., and Ikehara, M. (2021). High reflection removal using cnn with detection and estimation. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1381–1385.

Hinniger, C. and Rüter, J. (2023). Synthetic training data for semantic segmentation of the environment from uav perspective. *Aerospace*, 10(7).

Horyna, J., Krátký, V., Pritzl, V., Báča, T., Ferrante, E., and Saska, M. (2024). Fast swarming of uavs in gnss-denied feature-poor environments without explicit communication. *IEEE Robotics and Automation Letters*, 9(6):5284–5291.

Ji, D., Jin, W., Lu, H., and Zhao, F. (2024). Pptformer: Pseudo multi-perspective transformer for uav segmentation.

Li, H., Cai, Y., Hong, J., Xu, P., Cheng, H., Zhu, X., Hu, B., Hao, Z., and Fan, Z. (2023). Vg-swarm: A vision-based gene regulation network for uavs swarm behav-

ior emergence. *IEEE Robotics and Automation Letters*, 8(3):1175–1182.

Li, R. and Zhao, X. (2025). Aeroreformer: Aerial referring transformer for uav-based referring image segmentation.

Licea, D. B., Walter, V., Ghogho, M., and Saska, M. (2023). Optical communication-based identification for multi-uav systems: theory and practice.

Maxey, C., Choi, J., Lee, H., Manocha, D., and Kwon, H. (2024). Uav-sim: Nerf-based synthetic data generation for uav-based perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5323–5329.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2021). Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106.

Oh, X., Lim, R., Foong, S., and Tan, U.-X. (2023). Marker-based localization system using an active ptz camera and cnn-based ellipse detection. *IEEE/ASME Transactions on Mechatronics*, 28(4):1984–1992.

Pargieła, K. (2022). Vehicle detection and masking in uav images using yolo to improve photogrammetric products. *Reports on Geodesy and Geoinformatics*, 114:15–23.

Schilling, F., Schiano, F., and Floreano, D. (2021). Vision-based drone flocking in outdoor environments. *IEEE Robotics and Automation Letters*, 6(2):2954–2961.

Soltani, S., Ferlian, O., Eisenhauer, N., Feilhauer, H., and Kattenborn, T. (2024). From simple labels to semantic image segmentation: leveraging citizen science plant photographs for tree species mapping in drone imagery. *Biogeosciences*, 21:2909–2935.

Walter, V., Saska, M., and Franchi, A. (2018). Fast mutual relative localization of uavs using ultraviolet led markers. In *2018 International Conference on Unmanned Aircraft System (ICUAS 2018)*.

Walter, V., Staub, N., Franchi, A., and Saska, M. (2019). Uvdar system for visual relative localization with application to leader–follower formations of multirotor uavs. *IEEE Robotics and Automation Letters*, 4(3):2637–2644.

Xu, H., Wang, L., Zhang, Y., Qiu, K., and Shen, S. (2020). Decentralized visual-inertial-uwb fusion for relative state estimation of aerial swarm. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8776–8782.

Zhao, J., Li, Q., Chi, P., and Wang, Y. (2025). Active vision-based uav swarm with limited fov flocking in communication-denied scenarios. *IEEE Transactions on Instrumentation and Measurement*, pages 1–1.

Zhou, W., Zheng, N., and Wang, C. (2024). Synthesizing realistic traffic events from uav perspectives: A mask-guided generative approach based on style-modulated transformer. *IEEE Transactions on Intelligent Vehicles*, pages 1–16.

Zhou, X., Wen, X., Wang, Z., Gao, Y., Li, H., Wang, Q., Yang, T., Lu, H., Cao, Y., Xu, C., and Gao, F. (2022). Swarm of micro flying robots in the wild. *Science Robotics*, 7(66):eabm5954.