SegFly: A Dataset and 2D-3D-2D Paradigm for Aerial RGB-Thermal Semantic Segmentation at Scale

Markus Gross[1,2,3,⋆], Sai B. Matha[1], Rui Song[4,5], Viswanathan Muthuveerappan[1]
Conrad Christoph[1], Julius Huber[1], Daniel Cremers[2,3]

[1]Fraunhofer Institute IVI [2]TU Munich [3]MCML [4]UCLA [5]Uni Cambridge

Abstract. Semantic segmentation for uncrewed aerial vehicles (UAVs) is fundamental for aerial scene understanding, yet existing RGB and RGB-T datasets remain limited in scale, diversity, and annotation efficiency due to the high cost of manual labeling and the difficulties of accurate RGB-T alignment on off-the-shelf UAVs. To address these challenges, we propose a scalable geometry-driven 2D-3D-2D paradigm that leverages multi-view redundancy in high-overlap aerial imagery to automatically propagate labels from a small subset of manually annotated RGB images to both RGB and thermal modalities within a unified framework. By lifting less than 3% of RGB images into a semantic 3D point cloud and rendering it into all views, our approach enables dense pseudo ground-truth generation across large image collections, automatically producing 97% of RGB labels and 100% of thermal labels while achieving 91% and 88% annotation accuracy without any 2D manual refinement. We further extend this 2D–3D–2D paradigm to cross-modal image registration, using 3D geometry as an intermediate alignment space to obtain fully automatic, strong pixel-level RGB-T alignment with 87% registration accuracy and no hardware-level synchronization. Applying our framework to existing geo-referenced aerial imagery, we construct SegFly, a large-scale benchmark with over 20,000 high-resolution RGB images and more than 15,000 geometrically aligned RGB-T pairs spanning diverse urban, industrial, and rural environments across multiple altitudes and seasons. On SegFly, we establish the Firefly baseline for RGB and thermal semantic segmentation and show that both conventional architectures and vision foundation models benefit substantially from SegFly supervision, highlighting the potential of geometry-driven 2D-3D-2D pipelines for scalable multi-modal aerial scene understanding. The SegFly dataset and our Firefly baseline are available at https://github.com/markus-42/SegFly.

1 Introduction

Uncrewed aerial vehicles (UAVs) equipped with visual sensors have become a central platform for large-scale environment perception in computer vision.

⋆Corresponding author: markus.gross@tum.de.

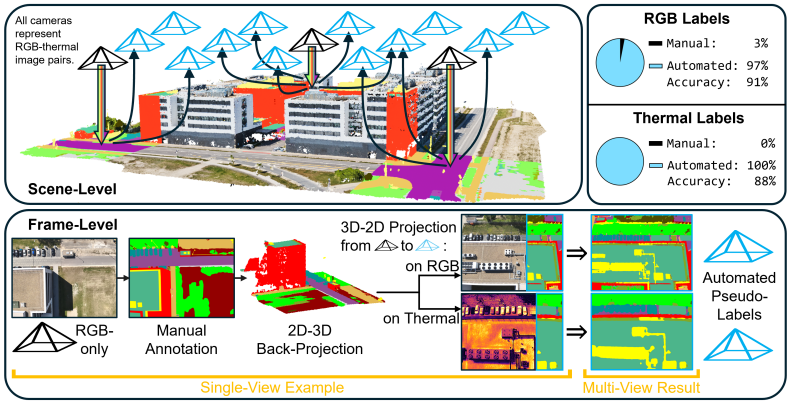

Fig. 1: 2D-3D-2D Paradigm. An overview is provided in Sec. 3.1.

Recent advances in aerial vision enable applications such as delivery and logistics, infrastructure inspection, environmental monitoring, emergency response, surveillance, precision agriculture, and urban planning [15, 16, 35]. These systems rely on robust scene understanding from aerial imagery captured under diverse environmental conditions. While most UAV platforms rely on RGB cameras, thermal sensors provide complementary information that is largely invariant to illumination changes and adverse weather, making RGB-Thermal (RGB-T) perception attractive for real-world deployments [25].

At the core of aerial perception are fundamental vision tasks including object detection [12, 27], tracking [41], localization [11], mapping and navigation [40], and 3D reconstruction [17]. Among these, semantic segmentation provides dense pixel-level scene understanding that directly supports higher-level reasoning and decision making [10]. However, large-scale aerial semantic segmentation remains challenging. Aerial imagery exhibits strong viewpoint variability, large scale changes, and complex scene layouts. For RGB-T perception, datasets additionally require accurate cross-modal alignment, which is difficult due to asynchronous capture, platform motion, and sensor characteristics. Existing solutions therefore rely on synchronized hardware or extensive manual refinement.

Another major limitation is the cost of dense pixel-level annotation. High-resolution aerial imagery contains complex structures that require careful labeling, making large-scale annotation expensive. As a result, most aerial segmentation datasets remain limited in scale and diversity, particularly for RGB-T data where aligned image pairs are required. Current datasets typically contain only a few thousand samples, restricting the ability of modern deep learning models to fully benefit from large-scale training data. Recent work has explored geometry-driven supervision to address these challenges. The 2D-to-3D label lifting paradigm leverages multi-view imagery to propagate semantic annotations through reconstructed geometry. The OccuFly benchmark [17] demonstrated that labels from a small subset of annotated images can be lifted into a 3D point cloud and reused

across multiple views. However, existing approaches focus on RGB imagery and do not address large-scale multi-modal segmentation or RGB-T alignment.

In this work, we present a scalable framework for aerial RGB and thermal semantic segmentation based on a geometry-driven 2D–3D–2D paradigm, shown in Fig. 1. By annotating only a small subset of RGB images, semantic labels are lifted into a reconstructed 3D point cloud and reprojected into both RGB and thermal views to generate dense pseudo ground truth at scale. Thermal imagery is additionally reconstructed in 3D and aligned with the RGB semantic point cloud, enabling accurate RGB-T registration through geometric alignment and lens distortion transfer entirely in software, eliminating the need for hardware synchronization and enabling scalable data collection with off-the-shelf UAV platforms. Building on this pipeline, we introduce SegFly[1], a large-scale benchmark for aerial semantic segmentation across RGB and RGB-T modalities, comprising more than 20,000 high-resolution RGB images and over 15,000 geometrically aligned RGB-T pairs captured across diverse environments, seasons, and flight altitudes. This scale is enabled by geometry-driven pseudo-label generation, requiring manual annotation for less than 3% of the collected data. Using SegFly, we establish baseline results with the Firefly architecture, providing a strong reference for both RGB and RGB-T segmentation. Furthermore, we show that RGB and RGB-T vision foundation models benefit substantially from training on SegFly, highlighting the effectiveness of large-scale geometry-derived supervision for multi-modal aerial perception.

In summary, our contributions are as follows:

- We propose a scalable **2D-3D-2D framework** that lifts manual semantic annotations from $<3\,\%$ of RGB images into a reconstructed 3D point cloud and renders dense pseudo ground-truth across $97\,\%$ of RGB and $100\,\%$ of thermal views, achieving $91\,\%$ and $88\,\%$ accuracy without human refinement.
- Based on our framework, we present **SegFly, a large-scale benchmark** for aerial semantic segmentation providing $>20,000$ RGB samples and $>15,000$ aligned RGB-T samples across diverse environments, seasons, and altitudes.
- Following the 2D–3D–2D paradigm, we propose a **multi-modal geometry-driven RGB-T registration method**, achieving $87.05\,\%$ pixel-level accuracy, thereby eliminating hardware synchronization. This enables practical and scalable data collection with off-the-shelf multi-modal UAVs, significantly reducing barriers to the construction of large-scale RGB-T datasets.
- On SegFly, we establish the **Firefly baseline**, a simple yet powerful and efficient method for aerial RGB and thermal semantic segmentation.
- We further show that RGB and RGB-T **aerial vision foundation models** benefit substantially from SegFly supervision, highlighting the potential to further reduce manual annotation toward fully automated data generation.

[1] SegFly dataset available at https://huggingface.co/datasets/markus-42/SegFly

Table 1: SegFly in comparison to established aerial RGB and RGB-T datasets.

Dataset	Modality	RGB-T Aligned	Resolution RGB	Resolution Thermal	Altitude [m]	# of Samples	# of Classes	Annotation Density	Environment Urban	Industrial	Rural	Venue
RGB Datasets												
Aeroscapes [28]	RGB	-	1280×720	-	5–30	3,269	12	Low	✓		✓	WACV '18
Okutswiss [39]	RGB	-	4608×3456 / 3840×2160	-	70; 80	191	10	Medium	✓		✓	JFR '22
ICG [20]	RGB	-	6000×4000	-	5–50	600	20	High	✓			-
UAVid [24]	RGB	-	4096×2160 / 3840×2160	-	50	300	8	Medium	✓			ISPRS JPRS '20
UDD6 [7]	RGB	-	4096×2160 / 4000×3000	-	60–100	141	6	Medium	✓			PRCV '18
VDD [3]	RGB	-	4000×3000	-	50–120	400	7	Medium	✓	✓	✓	JVCIR '25
Floodnet [32]	RGB	-	4000×3000	-	"low altitude"	2,343	10	Medium	✓			IEEE Access '21
UAVScenes [44]	RGB	-	2448×2048	-	80; 90; 130	120,000	19	Medium			✓	ICCV '25
MESSI [31]	RGB	-	5472×3684	-	30–100	2,525	16	High	✓			TMLR '25
CrossLoc [48]	RGB	-	720×480	-	-	7,000	7	High	✓		✓	CVPR '22
SkyScapes [1]	RGB	-	5616×3744	-	1,000	16	31	High	✓		✓	ICCV '19
SegFly-RGB (ours)	RGB	-	5472×3648 / 4000×3000	-	30; 40; 50	20,606	15	High	✓	✓	✓	
RGB-Thermal Datasets												
Kust4K [30]	RGB+T	✓	640×512	640×512	-	4,024	8	Medium	✓			Nature Sci-Data '25
CART [22]	RGB+T	✓	960×600	640×512	40	2,282	11	High			✓	ECCV '24
MVUAV [8]	RGB+T	✓	1920×1080	-	5–20	2,183	36	High	✓			NeurIPS '24
IndraEye [26]	RGB+T	✗	1280×720	640×480	7–30	5,612	13	Low	✓			CVPRW '25
SegFly-RGB-T (ours)	RGB+T	✓	640×512	640×512	30; 40; 50	15,007	15	High	✓	✓	✓	

2 Related Work

Datasets. Semantic segmentation from aerial platforms has a long history, which is reflected in established RGB datasets [1,3,7,24,28,31,32,39,44,48] and RGB-T datasets [8, 22, 26, 30], summarized in Tab. 1. Early works [7, 24, 28] provide high-resolution imagery with pixel-wise labels, but are limited to a few hundred images and a small number of classes, while more recent datasets [22,30,44,48] push toward higher class diversity and overall scale. Notably, RGB and RGB-T domains share the same bottleneck: they rely on manual 2D annotation as the primary source of supervision, resulting in costly and time-consuming human effort that limits scalability. To mitigate this bottleneck, existing work renders synthetic data [48], projects 3D labels with non-trivial manual refinement [44], or exploits foundation models to partially automate 2D masks [30]. However, existing approaches remain inefficient, as they still require substantial manual effort, including hundreds of man-hours, to maintain consistency across large numbers of small and thin structures [28,44]. Furthermore, for the RGB-T domain, existing methods inherit strong assumptions about data acquisition to ensure aligned RGB and thermal images that are necessary for improved segmentation results. Specifically, they rely on custom, hardware-synchronized sensor rigs with static extrinsics [8, 22, 25], or perform SIFT-based registration and manual cropping [30]. As a result, registration quality and scalability are tied to specific hardware configurations or substantial manual effort.

In contrast, SegFly is designed from the outset for off-the-shelf multi-modal UAVs that lack hardware-level time synchronization. By reconstructing RGB and thermal imagery in 3D, aligning their point clouds, and then transferring lens distortion back to 2D, our 2D–3D–2D pipeline achieves geometry-guided, precise RGB-T alignment entirely in software. Beyond RGB-T registration, we consider recent insights in 3D aerial semantic vision [17] to utilize our 2D-3D-2D approach for semantic labeling of both modalities: instead of annotating every

frame independently, we exploit the multi-view redundancy and metric accuracy of high-overlap, geo-referenced RGB-T imagery, by annotating a small subset of *RGB-only* images in 2D (<3 %), lift them into a semantic 3D point cloud, and render it back into the remaining >97 % of RGB views and 100 % of thermal views.

Methods. Apart from established deep semantic segmentation architectures such as UPerNet [46] and SegFormer [46], recent progress in the RGB domain has shifted toward leveraging vision foundation models [37] for generalizable dense prediction, often combined with parameter-efficient fine-tuning [6, 19, 45]. Vision-based segmentation foundation models, such as the SegmentAnything-Model [21] and adaptations such as SemanticSegmentAnything [4], provide broad generalization and interactive segmentation capabilities beyond fixed taxonomies. However, despite these advances, there remains tangible scope for improvement in aerial domains [29], where unique challenges such as extreme scale variation and viewpoint geometry degrade performance. To address these challenges in the aerial RGB domain, approaches like CatSeg [9] and SegEarth-OV [23] introduce open-vocabulary or cost-aggregation mechanisms to improve segmentation generality, while in the RGB-T domain, models such as AnyThermal [25] learn universal thermal representations via distillation from visual foundation models to better support thermal segmentation. We evaluate these models in a zero-shot setting and find that their performance on aerial imagery with the complexity present in SegFly is limited. However, when fine-tuned with SegFly, their performance improves substantially, highlighting the potential to further reduce manual annotation toward fully automated 2D and 3D aerial semantic data generation.

3 Data Generation Framework

3.1 Overview

Our data generation pipeline proposes a scalable *2D-3D-2D* paradigm that exploits the inherent multi-view redundancy of high-overlap aerial imagery to produce dense, high-quality semantic pseudo-labels for both RGB and thermal images at unprecedented scale. First, we perform classical multi-view geometry on the RGB images alone, yielding a metric 3D point cloud together with dense 2D–3D correspondences. In line with OccuFly [17], labels are then lifted from a small subset of manually annotated RGB images into this 3D point cloud, producing a semantically labeled point cloud with minimal human effort (see Sec. 3.3).

In the novel 3D–2D stage, we project the semantic point cloud onto *every* target frame, both the remaining RGB images and their corresponding thermal images. To address sparsity and occlusion artifacts inherent to point-based projection, we apply a four-step depth-guided semantic projection process: (1) semantic projection with Z-buffering, (2) depth-aware occlusion filtering, (3) splatting-based densification, and (4) depth-guided label propagation. The result is a dense, high-fidelity pseudo-label map for both RGB and thermal modalities, for which we annotate only 3 % of the RGB images manually while automatically generating 97 % of RGB and 100 % of thermal pseudo-labels.

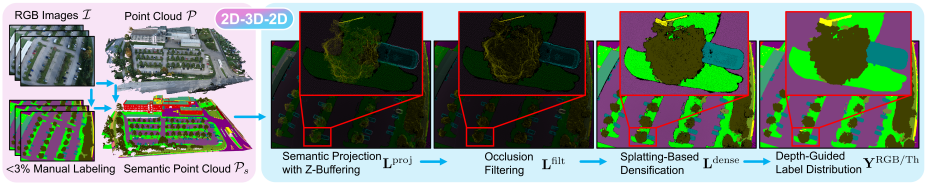

Fig. 2: Proposed 3D-to-2D point-based semantic rendering pipeline (Sec. 3.3).

Finally, to enable cross-modal domain adaptation, RGB-T image pairs must be registered. Rather than relying on cross-modal 2D feature matching or learnable homographies, we extend the 2D–3D–2D paradigm to cross-modal geometry-guided registration: thermal images are first reconstructed in 3D, aligned to the RGB point cloud using ICP, and then projected back to 2D via a lens-distortion transfer. This process achieves high-fidelity, strong pixel-level registration without hardware-level time-synchronization or strictly fixed extrinsics, enabling scalable data acquisition on off-the-shelf UAVs (see Sec. 3.4).

3.2 Problem Formulation

Let $\mathcal{I}^{\text{RGB}} = \{\mathbf{I}_n^{\text{RGB}} \in \mathbb{R}^{H\times W\times 3}\}_{n=1}^{N}$ and $\mathcal{I}^{\text{Th}} = \{\mathbf{I}_n^{\text{Th}} \in \mathbb{R}^{H\times W\times 1}\}_{n=1}^{N}$ denote the sets of RGB and thermal *image pairs*, respectively, acquired by a multi-modal UAV. All images have per-modality, geo-referenced world-to-camera poses $\mathcal{T}^{\text{RGB/Th}} = \{\mathbf{T}_n^{c\leftarrow w} \in \text{SE}(3)\}_{n=1}^{N}$ and intrinsics $\mathcal{K}^{\text{RGB/Th}} \in \mathbb{R}^{3\times 3}$. We define a semantic label set $\mathcal{C} = \{1,\dots,C\}$, and manual pixel-wise annotations are provided only for a small subset of RGB source images indexed by $\mathcal{S} \subset \{1,\dots,S\}$ with $|\mathcal{S}| \ll N$. The remaining images $\mathcal{U} = \{1,\dots,N\} \setminus \mathcal{S}$ are treated as target views. Our objective is to generate dense semantic annotations

$$\mathbf{Y}_n^{\text{RGB}},\ \mathbf{Y}_n^{\text{Th}} : \{1,\dots,H\} \times \{1,\dots,W\} \to \mathcal{C}, \quad \forall n \in \mathcal{U}, \tag{1}$$

for both RGB and thermal images, by exploiting a semantically labeled 3D point cloud $\mathcal{P}_{\mathcal{S}}$ reconstructed from the full RGB image set $\mathcal{I}^{\text{RGB}}$ and annotated only from the small subset of RGB source images $\mathcal{S}$.

Furthermore, in the context of cross-modal domain adaptation, for every thermal image $\mathcal{I}_n^{\text{Th}}$ we aim to register its corresponding RGB image $\mathcal{I}_n^{\text{RGB}}$, resulting in registered pairs $(\mathcal{I}_n^{\text{Th}}, \mathcal{I}_n^{\text{RGB,reg}})$.

3.3 2D–3D–2D Pseudo Ground-Truth Generation

2D-to-3D Stage: This stage directly follows the semantic 3D reconstruction pipeline of OccuFly [17], illustrated in Fig. 2. To summarize, using Structure-from-Motion (SfM) and Multi-View Stereo (MVS), we reconstruct a dense metric point cloud from all RGB images:

$$(\mathcal{P}, \boldsymbol{\kappa}^{\text{RGB}}, \mathcal{A}) = \Psi_{\text{SfM+MVS}}(\mathcal{I}^{\text{RGB}}, \mathcal{K}^{\text{RGB}}, \mathcal{T}^{\text{RGB}}), \tag{2}$$

where $\mathcal{P} = \{\mathbf{x}_m \in \mathbb{R}^3\}_{m=1}^{M}$ is the set of reconstructed 3D points, and $\boldsymbol{\kappa}^{\mathrm{RGB}}$ the estimated distortion coefficients. $\mathcal{A}_n$ are 2D–3D correspondences between a 3D point $\mathbf{x}$ and its 2D projection to pixel location (u, v) satisfying

$$((u, v), \mathbf{x}) \in \mathcal{A}_n \iff (u, v, 1)^\top \sim \mathbf{K}_n[\,\mathbf{R}_n \,|\, \mathbf{t}_n\,][\mathbf{x}^\top \; 1]^\top. \tag{3}$$

To minimize manual effort, we annotate only a small spatially stratified subset $\mathcal{S} \subset \{1, \ldots, S\}$ of the RGB images, selected by partitioning the ground plane into 25 m cells and choosing the closest center-pose per cell. This selection empirically ensures that over 99% of the reconstructed points are observed by at least one annotated image, while $|\mathcal{S}|/N < 0.1$ on average, quantified by the coverage metric

$$\frac{\left|\{\mathbf{x} \in \mathcal{P} \mid \exists n \in \mathcal{S},\, \exists (u, v) \text{ s.t. } ((u, v), \mathbf{x}) \in \mathcal{A}_n\}\right|}{|\mathcal{P}|} > 0.99. \tag{4}$$

After manual per-pixel annotation of images in $\mathcal{S}$, labels are lifted to 3D via back-projection using the correspondences $\mathcal{A}$. As most of the points are observed by multiple cameras, we fuse multi-view evidence by unweighted majority voting, where ties are broken by a fixed class-frequency prior. Unlabeled points are completed with inverse-distance-weighted kNN, followed by a second denoising kNN pass, analogous to [42]. The output is a semantically annotated point cloud

$$\mathcal{P}_{\mathcal{S}} = \{(\mathbf{x}_m, c_m)\}_{m=1}^{M}, \quad c_m \in \mathcal{C}. \tag{5}$$

3D-to-2D Stage: We now render the semantic point cloud $\mathcal{P}_{\mathcal{S}}$ into all target views $n \in \mathcal{U}$ to obtain dense pseudo-labels $\mathbf{Y}_n^{\mathrm{RGB/Th}}$ for both RGB and thermal images, illustrated in Fig. 2. Our rendering pipeline consists of four stages:

(1) Semantic Projection with Z-Buffering. For each target frame $n \in \mathcal{U}$ (RGB and thermal), we project every point $\mathbf{x}_m \in \mathcal{P}_{\mathcal{S}}$ into the image plane using the respective camera model. The projection includes full radial and tangential distortion correction, with RGB distortion coefficients $\boldsymbol{\kappa}^{\mathrm{RGB}}$ and thermal coefficients $\boldsymbol{\kappa}^{\mathrm{Th}}$ (discussed in Sec. 3.4). Moreover, we maintain a per-pixel Z-buffer to resolve depth conflicts, retaining only the point with smallest depth D at each pixel (u, v). This yields an initial sparse label map $\mathbf{L}_n^{\mathrm{proj}} \in (\mathcal{C} \cup \{0\})^{H \times W}$, where 0 denotes no projection.

(2) Occlusion Filtering. Projected points in the sparse label map $\mathbf{L}_n^{\mathrm{proj}}$ that survive Z-buffering may still be geometrically occluded by closer geometry of a different semantic class. We remove such inconsistent projections via a local kernel-based visibility filter. For each labeled pixel (u, v) with label c and depth $D(u, v)$, we consider a local neighborhood $\Omega_{u,v}$ of size $k \times k$. If there exists $(u', v') \in \Omega_{u,v}$ such that

$$\mathbf{L}_n^{\mathrm{proj}}(u', v') \neq c \quad \text{and} \quad D_n(u', v') < D_n(u, v) - \tau,$$

with depth threshold $\tau > 0$, we set $\mathbf{L}_n^{\mathrm{proj}}(u, v) = 0$ as likely occluded. We empirically determined $\tau = 0.2\,\mathrm{m}$, while $\Omega^{\mathrm{RGB}} = 9 \times 9$ and $\Omega^{\mathrm{Th}} = 5 \times 5$, due to

lower resolution of thermal images. This step transforms the initial sparse label $\mathbf{L}_n^{\text{proj}}$ to a filtered label map $\mathbf{L}_n^{\text{filt}}$, by removing geometrically inconsistent label projections.

(3) Splatting-based Densification. The filtered label map $\mathbf{L}_n^{\text{filt}}$ remains sparse due to point cloud sampling density. We increase density by splatting each labeled pixel to a small disk of fixed radius r_{splat}, increasing spatial coverage while preserving local consistency. To incorporate occlusion awareness, we process the splats in back-to-front order (decreasing depth), ensuring that nearer labels take precedence in overlapping regions, resulting in the densified label map $\mathbf{L}_n^{\text{dense}}$. In practice, $r_{\text{splat}}^{\text{RGB}} = 3$ and $r_{\text{splat}}^{\text{Th}} = 1$. The core of this step is standard point-splatting [49], extended to incorporate semantics labels and occlusion awareness.

(4) Depth-Guided Label Distribution. Remaining unlabeled pixels in $\mathbf{L}_n^{\text{dense}}$ are filled using a two-pass k-nearest-neighbor (kNN) scheme that explicitly respects depth ordering. In the first pass, for each unlabeled pixel, we retrieve its k nearest labeled neighbors in 2D image space. Among these, we assign the label of the neighbor with the smallest depth to the camera center, ensuring foreground priority. In a final pass, all pixels are refined using standard (depth-agnostic) 2D kNN with a larger neighborhood, ensuring consistent label boundaries. We empirically select $k = 5$ for the first pass, while for the second pass, k^{RGB} ranges from 7 to 25 and k^{Th} from 13 to 15. The resulting dense label maps $(\mathbf{Y}_n^{\text{RGB}}, \mathbf{Y}_n^{\text{Th}})$ constitute the pseudo ground-truth for their corresponding images $(\mathbf{I}_n^{\text{RGB}}, \mathbf{I}_n^{\text{Th}})$.

3.4 RGB-Thermal Image Registration

For every thermal image $\mathcal{I}_n^{\text{Th}}$ we aim to register its corresponding RGB image $\mathcal{I}_n^{\text{RGB}}$, resulting in registered pairs $(\mathcal{I}_n^{\text{Th}}, \mathcal{I}_n^{\text{RGB,reg}})$. To accomplish this, necessary RGB intrinsics, distortion coefficients, and poses $(\mathcal{K}^{\text{RGB}}, \boldsymbol{\kappa}^{\text{RGB}}, \mathcal{T}^{\text{RGB}})$ are obtained as described in Sec. 3.3. Thermal intrinsics and distortion coefficients $(\mathcal{K}^{\text{Th}}, \boldsymbol{\kappa}^{\text{Th}})$ are determined via a custom passive calibration target, following [36]. Similar to the RGB-based reconstruction in 3.3, we first perform an independent SfM and MVS reconstruction on the thermal images:

$$(\mathcal{P}^{\text{Th}}, \mathcal{T}^{\text{Th,ref}}) = \Psi_{\text{SfM+MVS}}(\mathcal{I}^{\text{Th}}, \mathcal{K}^{\text{Th}}, \mathcal{T}^{\text{Th}}), \tag{6}$$

yielding a thermal point cloud $\mathcal{P}^{\text{Th}}$ and globally refined thermal poses $\mathcal{T}^{\text{Th,ref}}$. RGB-T synchronization is then obtained by aligning $\mathcal{P}^{\text{Th}}$ to the RGB-based semantic point cloud $\mathcal{P}_{\mathcal{S}}$ using the Iterative Closest Point (ICP) algorithm [2], producing the rigid transformation $\mathbf{T}^{\text{reg}} \in \text{SE}(3)$. The registered thermal poses are given by

$$\mathcal{T}^{\text{Th,reg}} = \mathbf{T}^{\text{reg}} \circ \mathcal{T}^{\text{Th,rel}}, \tag{7}$$

where $\circ$ denotes the composition of rigid transformations in SE(3). This synchronization matches geometric structures across all images without requiring time-synchronization or static sensor rigs.

With registered thermal poses $\mathcal{T}^{\mathrm{Th,reg}}$ available, we formulate registration in a common undistorted pinhole domain shared by both cameras. Let $\bar{\pi}^{\mathrm{RGB}}$ and $\bar{\pi}^{\mathrm{Th}}$ denote the undistorted projection functions obtained by inverting the RGB and thermal distortion models $\boldsymbol{\kappa}^{\mathrm{RGB}}$ and $\boldsymbol{\kappa}^{\mathrm{Th}}$, respectively. For each thermal image n, we first determine the RGB support region corresponding to the thermal field of view.

To accomplish this, we first project all 3D points $\mathbf{x}_m \in \mathcal{P}_{\mathcal{S}}$ that are visible in the registered thermal camera into the undistorted thermal image using $\bar{\pi}^{\mathrm{Th}}$ and pose $\mathbf{T}_n^{\mathrm{Th,reg}}$, and then reproject the same points into the undistorted RGB image using $\bar{\pi}^{\mathrm{RGB}}$ and $\mathbf{T}_n^{\mathrm{RGB}}$. The extremal projected coordinates define an axis-aligned bounding box $\bar{\mathcal{B}}_n$ in this common pinhole domain. We then synthesize a registered RGB image in the thermal distortion space via a single lens-distortion transfer. Concretely, we resample the original (distorted) RGB image into the undistorted domain using the RGB camera warp (RGB intrinsics + $\boldsymbol{\kappa}^{\mathrm{RGB}}$) and subsequently map this undistorted RGB image into the thermal image plane using the thermal camera warp (thermal intrinsics + $\boldsymbol{\kappa}^{\mathrm{Th}}$). The same thermal camera warp is applied to the bounding box $\bar{\mathcal{B}}_n$, after which we crop and resize to the thermal resolution. This yields the registered RGB image $\mathbf{I}_n^{\mathrm{RGB,reg}}$, which is aligned with $\mathbf{I}_n^{\mathrm{Th}}$ in the thermal distortion space, closing the 2D–3D–2D loop for cross-modal registration.

By following the 2D-3D-2D paradigm, the resulting pairs $(\mathbf{I}_n^{\mathrm{Th}}, \mathbf{I}_n^{\mathrm{RGB,reg}})$ achieve highly accurate, strong pixel-level cross-modal consistency without cross-modal deterministic or learnable 2D features, homography estimation, manual correction, or hardware-level time synchronization, enabling high-fidelity registration on off-the-shelf UAVs, thereby streamlining data acquisition for scalable aerial RGB-T semantic segmentation.

4 Firefly Baseline

We introduce Firefly, a simple yet strong and efficient aerial baseline that addresses (1) RGB semantic segmentation, and (2) thermal semantic segmentation within a unified framework (see Fig. 3). Our objective is to achieve strong RGB performance while enabling accurate thermal segmentation through RGB domain adaptation that transfers knowledge from the RGB modality to the thermal domain in a parameter-efficient manner.

Training proceeds in three stages, termed RGB pre-training, RGB-Thermal domain adaptation, and thermal fine-tuning: In ***RGB pre-training***, a fixed DINOv3 [37] encoder is combined with a lightweight multi-layer perceptron (MLP) head [34] that is optimized on labeled RGB data to establish a strong baseline for RGB-based semantic segmentation. Subsequently, for ***RGB-Thermal domain adaptation***, registered RGB and thermal image pairs are leveraged to align feature representations in a self-supervised and parameter-efficient manner. This is accomplished by utilizing thermal-specific Rein [45] adapters, which learn residual mappings that transform thermal inputs into the RGB feature space, while keeping all other parameters fixed. Finally, for ***thermal fine-tuning***, a dedicated thermal

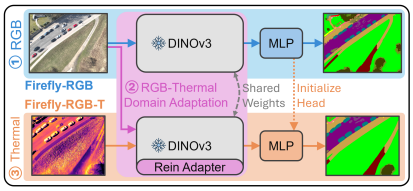

Fig. 3: Firefly overview (see Sec. 4).

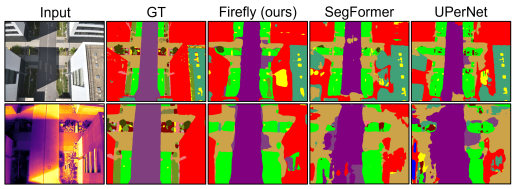

Fig. 4: Qualitative segmentation results (Sec. 6).

MLP head initialized from the RGB head is trained jointly with the Rein adapters on labeled thermal data, while the encoder and RGB head remain frozen, resulting in accurate thermal segmentation without compromising RGB performance. Full technical details of the adapter design, feature-level objectives, and optimization procedure are provided in the supplementary material.

5 SegFly Dataset

5.1 Raw Data

We build our dataset upon the geo-referenced aerial imagery of the OccuFly benchmark [17] (CVPR 2026), containing 9 distinct scenes. Data acquisition relies on two DJI UAV platforms: a Phantom 4 RTK (P4) [13] and a DJI Mavic 3 Enterprise Series (M3-ES) [14]. The P4 provides RGB imagery at a resolution of 5472×3648 pixels, whereas the M3-ES is the RGB–thermal-capable Mavic 3 Thermal configuration, providing 4000×3000 RGB frames and 640×512 long-wave infrared frames via software-based time-synchronization between both sensors. OccuFly scenes 1, 2, and 6-8 flown with the P4 thus yield RGB-only imagery, while scenes 3-5 and 9 flown with M3T produce paired RGB–T views. In total, our dataset leverages 20,606 geo-referenced RGB images and 15,007 geo-referenced thermal images with high multi-view overlap as the raw data source for 2D–3D–2D pseudo-label generation and RGB–T image registration. Finally, following OccuFly [17], we perform the 2D-to-3D label lifting per scene with OccuFly's total of 586 manually annotated RGB images captured at 50 m, but additionally exploit geo-referenced RGB and thermal imagery from 40 m and 30 m flights for the 3D-to-2D pseudo-labeling stage, effectively amortizing the fixed reconstruction and annotation cost at one altitude for a substantially larger image set without additional manual labeling.

5.2 Dataset Statistics

Overall statistics are summarized in Tab. 1. The dataset adheres to OccuFly's [17] data organization and comprises 9 scenes covering rural, urban, and industrial environments across all four seasons at 50 m, 40 m, and 30 m altitude. We present SegFly as an RGB-variant (***SegFly-RGB***) from OccuFly scenes 1-9, and an RGB-T-variant (***SegFly-RGB-T***) from OccuFly scenes 3, 4, 5, and 9. Importantly, we

Table 2: 2D-3D-2D label transfer quality and efficiency (see Sec. 5.3) by rendering a second set of manual GT that was *not* included during data generation.

Dataset	Modality	Altitude-wise Accuracy/FWmIoU [%] ↑				[%] of Manual Annotation ↓
		50m	40m	30m	combined	
SegFly-RGB (Ours)	RGB	91.30/85.90	91.74/85.57	91.38/86.10	**91.27/85.63**	**2.84**
SegFly-RGB-T (Ours)	Thermal	84.93/75.70	88.37/81.16	89.64/82.88	**87.65/79.46**	**0.00**

retain the original OccuFly [17] train/val/test split and augment it with additional RGB–T image pairs and 2D semantic annotations, effectively extending the underlying benchmark to a multi-modal 2D segmentation setting. Note that this implies that SegFly-RGB-T shares identical val/test splits, an intentional design choice to ensure sufficient sample counts and preserve the original organizational structure of the OccuFly dataset. Details regarding splits, seasons, environments, and altitude-wise sample counts are provided in the supplementary material. In summary, the RGB modality contains 20,606 samples, while the thermal modality comprises 15,007 samples. Every thermal sample forms an RGB–T pair, of which the RGB image was aligned to the thermal image by our registration pipeline described in Sec. 3.4. Finally, since our pseudo-labeling pipeline builds on OccuFly's classical SfM+MVS reconstruction under a static-scene assumption, dynamic objects or thin structures cannot always be reliably reconstructed in 3D. To prevent these ambiguous classes from propagating into the final pseudo-labels and to ensure high-quality supervision, we apply a semantic post-processing step in which the rare classes *Rock* (average class frequency 0.063 %) and *CableTower* (0.002 %) are merged into *GroundObstacle*, and the classes *Person*, *Bicycle*, *FlyingAnimals*, *Cables*, and *Cranes* (0.006 % combined) are mapped to an unlabeled category. Consequently, the final label set consists of 15 semantic classes, and 1.54/1.78 % of RGB/thermal pixels are unlabeled, primarily in images that cover the 3D scene boundaries during rendering. We retain these images to preserve realistic viewpoint statistics.

5.3 Dataset Evaluation

2D-3D-2D Label Transfer Accuracy and Efficiency. To ensure unbiased evaluation, we compare the generated pseudo-labels against *additional* manually annotated ground-truth that was *not* used during data generation (135 RGB images and 60 thermal images, distributed across all altitudes). We summarize labeling accuracy and efficiency in Tab. 2. Overall, the automatically generated RGB pseudo-labels achieve 91 % global pixel accuracy and 86 % class-frequency-weighted mean intersection over union (FWmIoU). Notably, this performance is achieved even though only 2.84 % of the RGB data was manually annotated during data generation, while the remaining 97.16 % was generated automatically. For thermal imagery, which is created entirely without manual annotation, we obtain 88/79 % accuracy/FWmIoU. Furthermore, the proposed multi-altitude strategy substantially improves labeling efficiency: manual annotation and semantic reconstruction are performed at a single altitude, and the results are rendered

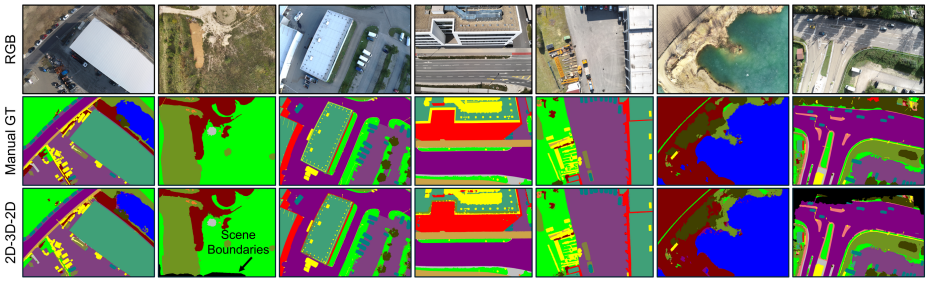

Fig. 5: Comparison of 2D-3D-2D pseudo-labels vs. manual ground-truth (Sec. 5.3).

into RGB and thermal images captured at multiple altitudes. Consequently, the fixed costs of reconstruction and manual labeling are amortized across a significantly larger number of images, thereby improving scalability without requiring additional human effort. Finally, Fig. 5 provides a qualitative comparison between manual ground truth and the generated pseudo-labels, highlighting the effectiveness of our label transfer. Despite the exceptionally high fidelity of OccuFly's [17] manual GT, the pseudo labels exhibit strong structural and semantic consistency, supporting the reliability of the proposed method.

3D-to-2D Rendering Evaluation. We ablate every major step of our proposed point-based semantic rendering pipeline by rendering and comparing with manual semantic ground-truth that was used for data generation, effectively synthesizing the ground-truth (see Tab. 5). This ablation shows that our proposed configuration yields the best results. We further compare our method with the rendering pipeline of UAVScenes [44] (ICCV 2025), which applies full manual 3D annotation followed by Screened Poisson Surface Reconstruction (SPSR). Although SPSR reconstructs a watertight mesh, the UAVScenes rendering procedure produces blended colors that do not match explicit semantic labels, yielding 64.07 % unlabeled pixels and an annotation accuracy/FWmIoU of only 34.92/34.48%. To mitigate these semantically ambiguous pixels, UAVScenes applies a second round of manual refinement, which undermines efficient and scalable data generation. In contrast, our pseudo-labels do not require manual refinement and achieve 90.08/85.14% annotation accuracy/FWmIoU. Taken together, these results validate our point-based semantic rendering pipeline and highlight its value for scalable, high-fidelity semantic data generation.

Semantic Consistency between Manual and Pseudo-Labeled Data. We verify that pseudo-labeling does not alter the underlying semantic distribution: per-class pixel frequencies computed from manual annotations and from our 2D–3D–2D pseudo-labels differ by only 1.06% on average. A detailed analysis and visualization of these distributions are provided in the supplementary material.

Table 3: RGB semantic segmentation results (see Sec. 6) on SegFly-RGB test set.

Method	Train Set: Manual GT	Train Set: SegFly-RGB	All Classes [%]: Acc.	All Classes [%]: FWmIoU
Eval on Manual GT				
UPerNet [46]	✓		52.08	52.46
		✓	**52.13**	**54.36**
SegFormer [47]	✓		50.05	48.50
		✓	**51.48**	**51.36**
Firefly (ours)	✓		47.81	45.81
		✓	**53.40**	**54.90**
Eval on SegFly-RGB				
UPerNet [46]	✓		47.72	43.72
		✓	**48.38**	**46.73**
SegFormer [47]	✓		46.02	41.74
		✓	**48.01**	**44.46**
Firefly (ours)	✓		45.72	42.34
		✓	**49.70**	**50.44**

Table 4: RGB-T semantic segmentation results (Sec. 6) on SegFly-RGBT test set.

Method	Acc.	FWmIoU
RGB-T Finetuning from Scratch		
UPerNet [46]	31.36	27.82
SegFormer [47]	<u>33.26</u>	<u>28.33</u>
Firefly-RGB (ours)	**37.48**	**35.20**
RGB Pretraining + RGB-T Finetuning		
UPerNet [46]	33.48	32.21
SegFormer [47]	<u>34.26</u>	<u>30.66</u>
Firefly-RGB (ours)	**37.67**	**37.38**
RGB Pretraining → RGB-T Adaptation → RGB-T Finetuning		
UPerNet [46] + Rein [45]	34.74	31.98
SegFormer [47] + Rein [45]	<u>35.75</u>	<u>33.48</u>
Firefly-RGB-T (ours)	**41.06**	**43.83**

RGB-T Registration Evaluation. To quantitatively validate the accuracy of the RGB-T registration, we utilize the second set of manual ground-truth that was *not* used for data generation (60 pairs across all scenes and altitudes). Specifically, we apply the identical projection, lens-distortion transfer, and cropping procedure (used to construct $\mathbf{I}_n^{\mathrm{RGB,reg}}$) to the RGB ground-truth of that set. The resulting aligned labels are compared with the corresponding thermal labels, providing a direct measure of semantic-level registration accuracy. Our method yields a strong RGB-T registration accuracy/FWmIoU of 87.05/77.65%. We further ablate our method by neglecting thermal lens distortion transfer, resulting in inferior performance of 84.08/73.63%. Finally, we qualitatively compare our registration results with established RGB-T aligned datasets (illustrated in the supplementary), and show that SegFly-RGB-T exhibits stronger registration performance. In summary, high quantitative performance and visual superiority confirm that our registration achieves precise alignment across both modalities.

6 Benchmark Experiments

RGB Semantic Segmentation. We benchmark SegFly-RGB for RGB semantic segmentation by training the established segmentation models UPerNet [46] and SegFormer [9] with their original architectures and recommended optimization settings, and by comparing them to our Firefly-RGB baseline under the same training protocol. We consider two supervision regimes, namely training on the limited manual ground-truth of OccuFly and training on the full SegFly-RGB pseudo-label set, and evaluate each model both on the manual ground-truth test split and on the SegFly-RGB test split. Results in Tab. 3 and Fig. 4 show that, when supervised with SegFly-RGB, Firefly-RGB achieves the highest accuracy and class-frequency-weighted mean intersection over union (FWmIoU) across both evaluation splits, outperforming UPerNet and SegFormer. We provide class-wise metrics in the supplementary material. More importantly, across all three architectures and both evaluation protocols, models trained on SegFly-RGB consistently outperform those trained only on manual ground-truth, despite

Table 5: Ablation on the 3D-to-2D pseudo-labels from our proposed point-based semantic rendering pipeline against manual ground-truth of test set scene 8 (Sec. 5.3).

Method	Occlusion Filtering	Splatting based Densification	Depth Guided Label Distribution	Accuracy [%]	FWmIoU [%]
(A)		✓	✓	89.78	84.58
(B)	✓		✓	89.54	84.35
(C)	✓	✓		83.95	80.49
(D) (ours)	✓	✓	✓	**90.08**	**85.14**

Table 6: Evaluation on SegFly's capability to improve Vision Foundation Models for aerial RGB and RGB-T semantic segmentation, by zero-shot evaluation and finetuning on SegFly (see Sec. 6).

Method	Eval Type	Eval Set	Acc.	FWmIoU
SegFly-RGB Semantic Segmentation				
CatSeg [9] (CVPR 2024)	Zero-Shot	CART [22]	4.32	15.03
	Finetuned SegFly-RGB	CART [22]	**11.35**	**17.45**
SegFly-RGB-T Semantic Segmentation				
AnyThermal [25] (ICRA 2026)	Zero-Shot	Kust4K [30]	46.02	35.06
	Finetuned SegFly-RGB-T	Kust4K [30]	**58.73**	**51.75**

the fact that SegFly-RGB pseudo-labels are generated automatically from a small fraction of human annotations. This demonstrates that our 2D-3D-2D pseudo-label pipeline provides a more effective supervisory signal than sparse manual labels alone and enables higher accuracy without additional human effort. Additionally, we ablate the MLP head of our Firefly baseline with ASPP [5] and DPT [33]. Results in the supplementary material show that our choice of MLP offers a more favorable trade-off between accuracy and model complexity.

Unsupervised RGB Semantic Segmentation. To verify that the improvements in Tab. 3 are not merely due to training on more RGB images, we perform an additional control experiment. Recall that the manual supervision regime uses only the 586 manually annotated SegFly-RGB frames ($\approx 3\%$ of the dataset), whereas the pseudo-label regime leverages all 20,606 SegFly-RGB images. In both cases, performance is evaluated on the same held-out test scenes, using either the manual annotations or the corresponding SegFly-RGB pseudo-labels. As a control, we train PriMaPs [18], a recent unsupervised segmentation method, on the same full set of SegFly-RGB images but without labels, thereby matching the number of training images while removing the supervisory signal from our pseudo-labels. PriMaPs achieves 31.19/16.84% and 31.64/15.31% Acc/FWmIoU on the manually annotated test set and the full SegFly-RGB test split, respectively. In contrast, our Firefly-RGB method improves these scores by $+71/+226\%$ and $+57/+229\%$ Acc/FWmIoU, respectively. These results demonstrate that the gains in Tab. 3 stem from the additional supervision provided by our 2D–3D–2D pseudo-label framework rather than a larger sample count.

Thermal Semantic Segmentation. We ablate Firefly's domain adaptation module Rein [45] with LoRA [19] and Adapter [6]. Detailed results presented in the supplementary show that our choices of MLP and Rein yield superior results. Moreover, we benchmark thermal semantic segmentation on SegFly-RGB-T by extending the three-stage Firefly training scheme (discussed in Sec. 4) to UPerNet [46] and SegFormer [47]. Results in Tab. 4 and Fig. 4 highlight that for all models, RGB pretraining consistently improves thermal performance over thermal-only training, and adding cross-modal adaptation further boosts accuracy

and FWmIoU for SegFormer and especially for Firefly, where Firefly-RGB-T achieves the best results with 41.06% accuracy and 43.83% FWmIoU. These findings show that SegFly-RGB and SegFly-RGB-T, together with the proposed RGB pretraining and RGB-T adaptation scheme, provide an effective setting for learning thermal semantic segmentation and for transferring knowledge from RGB to the thermal modality, most prominently for our Firefly baseline. We provide additional qualitative examples and full class-wise metrics in the supplementary.

Semantic Segmentation with RGB-Thermal Fusion. To assess the effect of RGB–thermal fusion, we extend our Firefly baseline to a multimodal variant on SegFly-RGB-T by concatenating RGB and thermal encoder features before the segmentation head and training under the same protocol as the thermal-only model. This fusion improves Firefly to 42.33/44.26% Acc/FWmIoU, corresponding to a $+1.27/+0.43\%$ gain over the thermal-only variant. These results suggest that RGB provides only marginal additional information beyond thermal under our training protocol, highlighting the strength of our RGB→thermal adaptation in effectively distilling RGB supervision into the thermal modality.

Vision Foundation Models. We evaluate how SegFly supports state-of-the-art vision foundation models for aerial RGB and thermal segmentation, presented in Tab. 6. We consider CatSeg [9] (CVPR 2024) and AnyThermal [25] (ICRA 2026) for RGB and thermal semantic segmentation, respectively. Following a cross-dataset evaluation, we first evaluate these pretrained models directly on an independent target dataset in a zero-shot manner. Subsequently, we fine-tune the models on SegFly and evaluate them again on the same target dataset. Keeping the evaluation set fixed allows us to quantify how fine-tuning on our data affects cross-dataset generalization. On SegFly-RGB, CatSeg improves from 4.32/15.03% accuracy/FWmIoU in zero-shot mode to 11.35/17.45% accuracy/FWmIoU after fine-tuning. On SegFly-RGB-T, AnyThermal improves from 46.02/35.06% in zero-shot mode to 58.73/51.75% accuracy/FWmIoU after fine-tuning. Together, these gains highlight that our dataset provides effective supervision that generalizes beyond the training domain, significantly improving cross-dataset generalization.

7 Conclusion

We present SegFly as an aerial RGB-T benchmark built with a 2D-3D-2D paradigm that turns a small set of manual RGB annotations into dense RGB and registered RGB-T labels across multiple altitudes. Experiments show that SegFly significantly improves RGB and thermal semantic segmentation for established architectures, recent aerial vision foundation models, and our proposed Firefly baseline. Future work will address remaining limitations, including dynamic objects by 4D methods [38], and the current restriction to daytime thermal imagery by investigating adversarial training strategies for robust adaptation across varying times of day [22, 43].

References

1. Azimi, S.M., Henry, C., Sommer, L., Schumann, A., Vig, E.: Skyscapes fine-grained semantic understanding of aerial scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7393–7403 (2019)
2. Besl, P.J., McKay, N.D.: A method for registration of 3-d shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **14**(2), 239–256 (1992). https://doi.org/10.1109/34.121791
3. Cai, W., Jin, K., Hou, J., Guo, C., Wu, L., Yang, W.: Vdd: Varied drone dataset for semantic segmentation. Journal of Visual Communication and Image Representation **109**, 104429 (2023)
4. Chen, J., Yang, Z., Zhang, L.: Semantic segment anything. https://github.com/fudan-zvg/Semantic-Segment-Anything (2023)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(4), 834–848 (2018)
6. Chen, S., GE, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 16664–16678. Curran Associates, Inc. (2022)
7. Chen, Y., Wang, Y., Lu, P., Chen, Y., Wang, G.: Large-scale structure from motion with semantic constraints of aerial images. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 347–359. Springer (2018)
8. Cheng, L., Ji, W., Jin, H., Li, J., Li, W., Shen, Y.: Unleashing multispectral video's potential in semantic segmentation: A semi-supervised viewpoint and new UAV-view benchmark. In: Advances in Neural Information Processing Systems 37. pp. 65717–65737. Neural Information Processing Systems Foundation, Inc. (NeurIPS), San Diego, California, USA (2024)
9. Cho, S., Shin, H., Hong, S., An, S., Lee, S., Arnab, A., Seo, P.H., Kim, S.: Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25357–25367 (June 2024)
10. Csurka, G., Volpi, R., Chidlovskii, B.: Semantic image segmentation: Two decades of research. Foundations and Trends in Computer Graphics and Vision **14**(1-2), 1–162 (2022). https://doi.org/10.1561/0600000095
11. Dhaouadi, O., Marin, R., Meier, J.M., Kaiser, J., Cremers, D.: Ortholoc: UAV 6-dof localization and calibration using orthographic geodata. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
12. Dhaouadi, O., Meier, J., Wahl, L., Kaiser, J., Scalerandi, L., Wandelburg, N., Zhuo, Z., Berinpanathan, N., Banzhaf, H., Cremers, D.: Highly accurate and diverse traffic data: The deepscenario open 3d dataset. In: 2025 IEEE Intelligent Vehicles Symposium. IEEE (2025)
13. DJI: Phantom 4 rtk. https://www.dji.com/phantom-4-rtk/info (2016), accessed: 2026-02-20
14. DJI: Mavic 3 enterprise series. https://enterprise.dji.com/mavic-3-enterprise (2022), accessed: 2026-02-20

15. Farajijalal, M., Eslamiat, H., Avineni, V., Hettel, E., Lindsay, C.: Safety systems for emergency landing of civilian unmanned aerial vehicles (uavs)—a comprehensive review. Drones **9**(2) (2025), https://doi.org/10.3390/drones9020141
16. Gross, M., Greiner, A., Matha, S.B., Soest, F., Cremers, D., Meeß, H.: SafeLand: Safe Autonomous Landing in Unknown Environments with Bayesian Semantic Mapping (2026), https://arxiv.org/abs/2603.17430
17. Gross, M., Matha, S.B., Fahmy, A., Song, R., Cremers, D., Meess, H.: OccuFly: A 3D Vision Benchmark for Semantic Scene Completion from the Aerial Perspective (2025), https://arxiv.org/abs/2512.20770
18. Hahn, O., Araslanov, N., Schaub-Meyer, S., Roth, S.: Boosting unsupervised semantic segmentation with principal mask proposals. Transactions on Machine Learning Research (TMLR) (2024)
19. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
20. Institute of Computer Graphics and Vision, Graz University of Technology: Semantic drone dataset (2019)
21. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023)
22. Lee, C., et al.: Caltech aerial rgb-thermal dataset in the wild. In: European Conference on Computer Vision (ECCV). Springer (2024)
23. Li, K., Liu, R., Cao, X., Bai, X., Zhou, F., Meng, D., Wang, Z.: Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10545–10556 (June 2025)
24. Lyu, Y., Vosselman, G., Xia, G.S., Yilmaz, A., Yang, M.Y.: Uavid: A semantic segmentation dataset for uav imagery. ISPRS journal of photogrammetry and remote sensing **165**, 108–119 (2020)
25. Maheshwari, P., Karhade, J., Chawla, Y., Adu, I., Heisen, F., Porco, A., Jong, A., Liu, Y., Pitla, S., Scherer, S., Wang, W.: Anythermal: Towards learning universal representations for thermal perception (2026), https://arxiv.org/abs/2602.06203
26. Manjunath, D., Sikdar, A., Gurunath, P., Udupa, S., Sundaram, S.: Saga: Semantic-aware gray color augmentation for visible-to-thermal domain adaptation across multi-view drone and ground-based vision systems. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 4578–4588. IEEE (2025)
27. Meier, J., Scalerandi, L., Dhaouadi, O., Kaiser, J., Araslanov, N., Cremers, D.: Carla drone: Monocular 3d object detection from a different perspective. In: Pattern Recognition - 46th DAGM GCPR 2024. pp. 137–152. Lecture Notes in Computer Science, Springer Science and Business Media Deutschland GmbH (2025). https://doi.org/10.1007/978-3-031-85187-2_9
28. Nigam, I., Huang, C., Ramanan, D.: Ensemble knowledge transfer for semantic segmentation. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1499–1508. IEEE (2018)
29. Osco, L.P., Wu, Q., de Lemos, E.L., Gonçalves, W.N., Ramos, A.P.M., Li, J., Marcato, J.: The segment anything model (sam) for remote sensing applications: From

zero to one shot. International Journal of Applied Earth Observation and Geoinformation **124**, 103540 (2023). https://doi.org/https://doi.org/10.1016/j.jag.2023.103540
30. OuYang, J., Wang, Q., Shang, Y., Jin, P., Zhong, H., Zhou, L., Shen, T.: An rgb-tir dataset from uav platform for robust urban traffic scenes semantic segmentation. Sci Data **12**, 1701 (2025). https://doi.org/10.1038/s41597-025-05994-7
31. Pinkovich, B., Matalon, B., Rivlin, E., Rotstein, H.: MESSI: A multi-elevation semantic segmentation image dataset of an urban environment. Transactions on Machine Learning Research (2025)
32. Rahnemoonfar, M., Chowdhury, T., Sarkar, A., Varshney, D., Yari, M., Murphy, R.: Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. IEEE Access **9**, 89644–89654 (2021), publisher Copyright: © 2013 IEEE.
33. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12179–12188 (October 2021)
34. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986)
35. Semerikov, S.O., Nechypurenko, P.P., Vakaliuk, T.A., Mintii, I.S., Kolhatin, A.O.: Vision-based autonomous UAV landing: A comprehensive review of technologies, techniques, and applications. Journal of Intelligent and Robotic Systems **111**(4) (Oct 2025)
36. Shivakumar, S.S., Rodrigues, N., Zhou, A., Miller, I.D., Kumar, V., Taylor, C.J.: Pst900: Rgb-thermal calibration, dataset and segmentation network. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 9662–9669. IEEE (2020)
37. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025), https://arxiv.org/abs/2508.10104
38. Song, R., Liang, C., Xia, Y., Zimmer, W., Cao, H., Caesar, H., Festag, A., Knoll, A.: Coda-4dgs: Dynamic gaussian splatting with context and deformation awareness for autonomous driving. In: IEEE/CVF International Conference on Computer Vision (ICCV). IEEE/CVF (2025)
39. Speth, S., Gonçalves, A., Rigault, B., Suzuki, S., Bouazizi, M., Matsuo, Y., Prendinger, H.: Deep learning with RGB and thermal images onboard a drone for monitoring operations. Journal of Field Robotics **39**(6), 840–868 (Sep 2022)
40. Stumberg, L., Usenko, V., Cremers, D.: Autonomous exploration with a low-cost quadrocopter using semi-dense monocular slam. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5775–5782. IEEE (2016)
41. Sturm, J., Bylow, E., Kerl, C., Kahl, F., Cremers, D.: Dense tracking and mapping with a quadrocopter. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences pp. 395–400 (2013)
42. Tian, X., Jiang, T., Yun, L., Wang, Y., Wang, Y., Zhao, H.: Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. arXiv preprint arXiv:2304.14365 (2023)
43. Vertens, J., Zürn, J., Burgard, W.: Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8461–8468 (2020)

44. Wang, S., Li, S., Zhang, Y., Yu, S., Yuan, S., She, R., Guo, Q., Zheng, J., Howe, O.K., Chandra, L., et al.: Uavscenes: A multi-modal dataset for uavs. arXiv preprint arXiv:2507.22412 (2025)
45. Wei, Z., Chen, L., Jin, Y., Ma, X., Liu, T., Ling, P., Wang, B., Chen, H., Zheng, J.: Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28619–28630 (June 2024)
46. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018)
47. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems. vol. 34, pp. 12077–12090 (2021)
48. Yan, Q., Zheng, J., Reding, S., Li, S., Doytchinov, I.: Crossloc: Scalable aerial localization assisted by multimodal synthetic data. arXiv preprint arXiv:2112.09081 (2021)
49. Zwicker, M., Pfister, H., Baar, J.v., Gross, M.: Surface splatting. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. pp. 371–378. SIGGRAPH '01, ACM, New York, NY, USA (2001)