POLITECNICO DI TORINO
Repository ISTITUZIONALE

Towards Robust DEM-Based Monocular Depth Rescaling for UAVs: A Systematic Analysis

Original
Towards Robust DEM-Based Monocular Depth Rescaling for UAVs: A Systematic Analysis / Musio, M.G., Savian, S., Mohammadi, S.S., Capello, E., Primatesta, S.. - ELETTRONICO. - (2026), pp. 167-174. (2026 International Conference on Unmanned Aircraft Systems, ICUAS 2026 Corfu (GRC) 15-18 June 2026) [10.1109/icuas69441.2026.11598587].

Availability:
This version is available at: 11583/3016614 since: 2026-11-02T12:47:28Z

Publisher:
Institute of Electrical and Electronics Engineers

Published
DOI:10.1109/icuas69441.2026.11598587

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

10 November 2026

Towards Robust DEM-Based Monocular Depth Rescaling for UAVs: A Systematic Analysis

Maria Grazia Musio[1], Stefano Savian[2], Seyed Saber Mohammadi[2], Elisa Capello[1], Stefano Primatesta[1]

***Abstract*— Monocular depth estimation suffers from inherent scale ambiguity, limiting its real-world applicability. While recent foundation models produce high-quality relative depth maps, their predictions lack metric consistency, which becomes critical in aerial environments. Recent rescaling approaches address this limitation by leveraging Digital Elevation Models (DEMs) as absolute references, yet existing DEM-based methods remain difficult to deploy due to their reliance on offline ground-segmentation techniques and the lack of standardized validation frameworks. To overcome these challenges, we introduce a Synthetic-to-Geo-Real simulation framework that jointly models realistic flight dynamics and real georeferenced terrain, enabling a novel, systematic, and controlled evaluation of DEM-based rescaling across diverse operational conditions. The framework provides a unified benchmarking setting to compare multiple state-of-the-art depth models and alternative ground segmentation strategies under consistent assumptions. Our results show that semantic segmentation provides more consistent performance across models, altitudes, and environments, achieving up to a 60% reduction in error over geometric segmentation. We further observe that decoupling segmentation from depth estimation and selectively sampling ground points can improve scale recovery. These findings support the practical viability of DEM-based metric depth rescaling for Unmanned Aerial Navigation.**

Keywords— Monocular Metric Depth Estimation, UAV, Ground Segmentation, Semantic Segmentation, Digital Elevation Maps, Scale Recovery

I. Introduction

The integration of Unmanned Aerial Vehicles (UAVs) with intelligent perception enables autonomous systems capable of interpreting their environment with minimal human intervention [1]. Monocular Depth Estimation (MDE) plays a central role in this context because of its low payload, power efficiency, and compatibility with embedded platforms. However, it suffers from scale ambiguity: multiple 3D points along the same camera ray project onto a single 2D pixel, so their absolute distances cannot be determined uniquely from a single view.

MDE has evolved from early lightweight architectures to billion-parameter foundation models [2]. Despite their remarkable zero-shot generalization capabilities, these models still produce only relative or incorrect metric depth maps, making explicit metric scale recovery essential for UAV deployment. Moreover, UAV operations introduce additional challenges that further reduce depth accuracy, including motion blur, illumination variability, and oblique viewing angles. Addressing these alongside scale ambiguity, current recovery methods face several limitations: Language-based approaches rely on object-scene priors [3], which are often missing in aerial imagery; geometry-based methods [4] are computationally intensive and prone to drift; sensor fusion with LiDAR [5] or other sensors increases payload and power consumption, critical constraints for small UAVs.

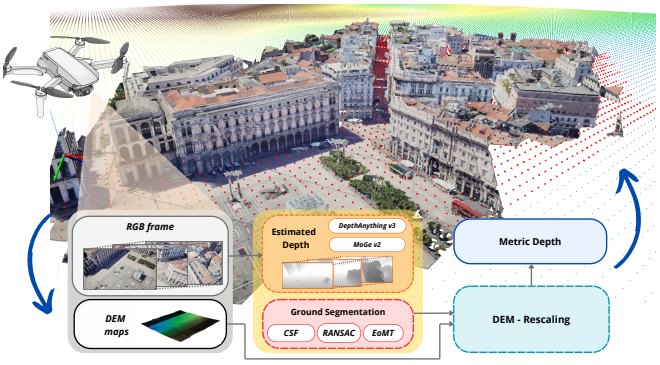


Figure 1: Overview of the proposed metric depth estimation pipeline for UAVs. Monocular depth models process RGB images; ground segmentation identifies terrain points, which are aligned with DEM elevations to recover metric scale. *Base imagery: © 2025 Google Maps.*

To overcome these challenges, TanDepth [6] addresses the scale ambiguity problem by aligning depth predictions with Digital Elevation Maps (DEMs). DEMs are widely available, freely accessible, and provide absolute metric references without additional sensors or calibration procedures. However, TanDepth's reliance on the Cloth Simulation Filter (CSF) [7] for ground segmentation introduces significant barriers to operational deployment. Originally designed for post-processing airborne LiDAR data with a known metric scale, the CSF struggles when applied to predicted depth with an unknown scale. This leads to three key issues: *(i)* high computational cost due to the iterative cloth simulation; *(ii)* strong hyperparameter sensitivity requiring scene-specific tuning; *(iii)* tight coupling between depth prediction characteristics and segmentation performance. As a result, CSF parameters cannot be adapted online to the varying depth distributions encountered along UAV trajectories, creating a critical bottleneck for robust autonomous operation. Moreover, the dataset used for evaluation in TanDepth is not publicly available, further limiting systematic assessment. Therefore, this work addresses the following research ques-

[1]M. G. Musio, E. Capello, and S. Primatesta are with the Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Torino, Italy. Corresponding author: S. Primatesta (e-mail: stefano.primatesta@polito.it)

[2]S. Savian, and S. S. Mohammadi are with Leonardo S.p.a - Center of Excellence (CoE), Genoa, Italy.

tion: *Can DEM-based monocular depth rescaling methods be redesigned to be robust and deployable in real-world aerial environments?*

Evaluating DEM-based scale recovery is challenging, as it requires both pixel-accurate depth ground truth and precise alignment with operational DEMs. Real-world UAV datasets [8]–[10] provide realistic imagery, but depth measurements are often noisy and sparse. Global Navigation Satellite System (GNSS) information is coarse, and datasets often feature uncalibrated magnetometers and inertial sensors, impeding an accurate DEM alignment. Synthetic datasets [11]–[13], in contrast, offer simulation-quality depth but depict virtual environments that lack real terrain orography, preventing direct evaluation of operational DEM-based methods such as TanDepth. To address this gap, we introduce a Synthetic-to-Geo-Real framework that integrates photorealistic simulation, realistic UAV flight dynamics, and real-world DEMs. By anchoring synthetic scenes to the real terrain, our framework enables a rigorous evaluation of DEM-based rescaling under realistic flight conditions while preserving the geometric fidelity required for accurate scale recovery. Furthermore, explicitly modeling real-world data allows direct comparison with real DEMs, providing deeper insight into operational challenges. Figure 1 illustrates the general DEM-based pipeline for metric scale recovery from monocular depth estimation.

The main contributions of this work are as follows:

- **Synthetic-to-Geo-Real evaluation methodology:** a novel framework combining photorealistic simulation, UAV dynamics, and real-world georeferenced DEMs for reproducible assessment of DEM-rescaling methods.
- **Unified framework for systematic assessment of DEM rescaling strategies:** a coherent benchmarking pipeline integrating multiple MDE models and ground segmentation approaches within a common evaluation setting, enabling fair and principled comparison between geometric and learning-based rescaling methods.
- **Ground quality versus extent analysis:** examining how the accuracy and spatial coverage of detected ground regions influence metric scale estimation.
- **Operational insights:** practical guidelines for integrating DEM-based scale recovery into UAV pipelines.

II. Related Work

A. Monocular Depth Estimation

MDE is an inherently ill-posed problem, resulting in metric scale ambiguity. Over the years, methods have evolved from supervised approaches [14], [15] to self-supervised [16], [17] and foundation models [18], [19]. Recent models, such as MoGe v2 [20], leverage affine-invariant point maps with robust alignment to achieve state-of-the-art relative geometry. At the same time, DepthAnything (DA) v3 [21] employs a compact transformer and teacher–student training to enable zero-shot cross-domain inference. Despite these advances, MDE remains particularly challenging for aerial viewpoints, due to oblique angles, varying altitudes, and distribution shifts relative to terrestrial training data [22].

B. Scale Recovery

Scale recovery methods differ in sensor requirements and assumptions. Sensor fusion [5], [23] increases payload; motion-based approaches [24], [25] require sufficient baselines and calibration; learning-based methods [19], [26] struggle with aerial generalization. TanDepth [6] introduces reference-based rescaling using global DEMs, but it remains unsuitable for real-time operational deployment.

C. Ground Segmentation Methods

Most methods for ground segmentation in LiDAR navigation [27] are well-established. These include geometric approaches such as RANSAC [28] and region growing [29], which struggle on irregular terrain. Morphological filters like CSF [7] and PMF [30] are sensitive to parameters and scene configurations. Learning-based methods [31], [32] offer greater robustness when trained on aerial data. Despite extensive research on MDE for UAVs [22] and LiDAR ground segmentation [27], the role of ground segmentation in MDE remains underexplored.

III. Methodology

A. Problem Formulation

Given an RGB image I, a monocular depth estimation model f_θ produces a predicted depth map $\hat{\mathbf{D}} = f_\theta(I)$ ideally related to the metric depth map $\mathbf{D}_{\text{metric}}$ through an unknown affine transformation where s scales and t shifts the predictions to match the metric depths.

$$\mathbf{D}_{\text{metric}} = s\,\hat{\mathbf{D}} + t. \tag{1}$$

For *relative depth models*, (s, t) are unknown and must be estimated. For *metric depth models* applied to aerial scenes, these parameters can still be biased due to differences between training and deployment data, and thus often require correction even though the models should output metric depths. DEM-rescaling estimates these parameters by identifying ground pixels $\mathcal{G} = \{(i, j)\}$ and aligning their predicted depths with DEM elevations d_{ij}^{DEM}:

$$(s, t) = \arg\min_{s,t} \sum_{(i,j)\in\mathcal{G}} \left(s\,\hat{d}_{ij} + t - d_{ij}^{\text{DEM}} \right)^2. \tag{2}$$

The reliability of (s, t) depends primarily on:

- **Ground pixel selection:** only true ground points should contribute to the affine alignment. Including non-ground points introduces mismatches between $\hat{d}_{ij}$ and d_{ij}^{DEM}, causing errors in scale and offset estimation.
- **Geometric consistency of the depth map:** Rescaling is reliable only if the depth map preserves internal geometry, as DEM-rescaling corrects global scale and offset but not local distortions or depth order errors.

B. Evaluation Protocol

As outlined in III-A, the DEM-rescaling pipeline has two core components: MDE and ground segmentation, whose interaction determines metric scale recovery quality. Our evaluation disentangles their contributions through separate protocols, using two complementary datasets. SkyScenes [33] provides dense depth and semantic annotations across controlled altitudes and camera angles, enabling statistical validation under controlled conditions. The Synthetic-to-Geo-Real dataset, evaluated at 70 m, anchors synthetic scenes to georeferenced urban terrain, verifying that the observed patterns generalize to operational environments with real DEM correspondence. The analysis is organized into the following three protocols:

Estimated Depth Evaluation: We evaluate two state-of-the-art MDE models targeting three aspects directly relevant to DEM-based rescaling:

- Absolute depth range characterization: analysis of the raw depth value distribution as a function of flight altitude (i.e., the UAV depth range), establishing the baseline against which rescaling effectiveness is evaluated.
- Relative depth consistency: evaluation of how well relative depth relationships are preserved, independently of the absolute scale.
- Error under ideal rescaling: error after optimal global affine transformation, using the Ground-Truth (GT) depth as the absolute reference. This bypasses DEM alignment and ground segmentation, isolating residual errors that cannot be corrected by global rescaling and defining the lower bound on the achievable error for each model.

Ground Segmentation Evaluation: Baseline segmentation performance is first evaluated on *SkyScenes*. Geometric methods are applied to both GT and model-predicted depth, enabling a clear separation between segmentation quality and sensitivity to depth-prediction errors, whereas the semantic method [31] is evaluated directly on RGB images. In the *Synthetic-to-Geo-Real* setting, dense GT annotations are unavailable. Therefore, three representative frames are manually annotated. These frames cover different levels of scene complexity, ranging from scenes with large ground regions to scenes with limited ground visibility.

DEM Rescaling Evaluation – Combined Effects of Depth and Segmentation: Each combination of depth model and ground segmentation method is tested along the Synthetic-to-Geo-Real trajectory. This captures how errors in depth prediction and segmentation propagate through affine rescaling and affect the final metric depth accuracy. For each configuration, the following analyses are reported:

- Trajectory-averaged depth metrics: Overall assessment of metric depth accuracy across the entire flight.
- Frame-level analysis: Identification of scene-dependent behavior, outliers, and failure cases.
- Altitude robustness: Validation across five altitudes (50, 70, 100, 120, 170 m) to assess generalization to operational height variations. This captures altitude-dependent challenges: changing ground visibility, depth range expansion, and geometric distortions that vary with UAV elevation.

C. Impact of Ground Spatial Sampling on DEM Rescaling

This section investigates how the amount and spatial distribution of ground pixels influence rescaling accuracy, using the best-performing depth-segmentation configuration identified in the previous section. The full ground mask is divided into a 4×4 spatial grid of 16 subregions, defined by rows r0–r3 and columns c0–c3. Each subregion is analyzed independently by estimating the affine transformation parameters through Equation (2). This enables a direct comparison between partial ground coverage and the full-ground baseline across operational altitudes, thereby quantifying the trade-off between ground coverage and rescaling robustness.

D. Computational Efficiency Evaluation

The high computational demand is another limitation of the CSF, directly affecting the deployability on embedded UAV platforms. Processing latency is quantified by decomposing the pipeline into three stages: depth network inference (t_{depth}), ground segmentation (t_{seg}), and affine rescaling computation via least-squares solve (t_{scale}). Experiments are conducted on a ground station equipped with an NVIDIA RTX A3000 GPU (12GB VRAM), not onboard a UAV. We report the mean latency over trajectory frames, isolating the dominant bottleneck and quantifying the speedup potential of alternative segmentation methods.

IV. Experimental Setup

This section describes the experimental framework used to evaluate the DEM-rescaling method performance. We first outline the selected MDE and ground segmentation methods, then describe the datasets, the Synthetic-to-Geo-Real environment, and flight mission configurations.

A. Depth Estimation Models

We evaluate two state-of-the-art foundation models, **DA v3** and **MoGe v2**, chosen for their strong benchmark performance and differing prediction strategies. Depth prediction accuracy is assessed using standard metrics: **Mean Absolute Error (MAE)** for absolute depth accuracy, **Absolute Relative Error (ABS Rel)** for scale-normalized error, and $\boldsymbol{\delta_{1.25}}$ for depth accuracy within a 1.25× threshold.

B. Ground Segmentation Methods

Three representative segmentation approaches are assessed:

- **CSF:** A virtual cloth is draped over an inverted point cloud to approximate terrain; points near the cloth are classified as ground. This matches the ground segmentation strategy of TanDepth, and its evaluation in our framework provides an approximate reimplementation of that baseline.
- **Multi-plane RANSAC:** Multiple planes are iteratively fitted, selecting the plane with the lowest average height

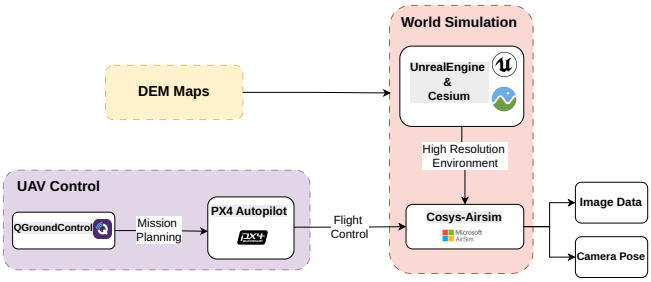


Figure 2: Synthetic-to-Geo-Real framework: UAV flights are simulated in a georeferenced Unreal Engine environment, generating RGB and depth data for DEM-based evaluation.

as ground. The point cloud is downsampled to reduce computational cost.

- **EoMT:** Ground regions are predicted from RGB using EoMT-Large (ViT-L/14 DINOv2, ADE20K panoptic segmentation). The ground mask is defined as the union of nine ADE20K classes (*floor, road, grass, sidewalk, earth, field, path, dirt track, land*) via per-pixel argmax, with no post-processing.

For both CSF and RANSAC, the parameters are manually tuned on a single dataset image to achieve optimal performance and then kept constant for all other images.

Segmentation quality is measured using standard metrics: **IoU** (Intersection over Union), **Precision**, **Recall**, and **F1-score**.

C. Datasets and Evaluation Environments

SkyScenes: SkyScenes is a synthetic UAV dataset generated with CARLA [34], providing pixel-accurate depth and dense semantic annotations. It covers urban and rural environments at altitudes of 15–60 m and camera pitch angles from 0° to 90°. To simulate real sensor behavior, nominal flight altitudes are perturbed with uniform noise. We use the ClearNoon subset for controlled evaluation under optimal visibility.

Synthetic-to-Geo-Real Framework: We introduce a simulation environment integrating photorealistic rendering with real-world georeferenced terrain to enable DEM-anchored validation under realistic flight conditions. The main architectural components are shown in Figure 2 and consist of:

a) Geospatial Foundation: Cesium [35] streams Google's photorealistic 3D Tiles [36] into Unreal Engine 5, providing a high-accuracy WGS84 globe with textured meshes (© Google Maps), aligned to the Tinitaly DEM [37] for precise terrain representation. Tinitaly provides 10 m resolution national coverage of Italy, derived from topographic maps, GNSS points, and radar altimetry, with a mean vertical accuracy below 3.5 m [38].

b) UAV Simulation: Cosys-AirSim [39] provides physics-based multirotor dynamics with PX4 autopilot [40]. The simulated UAV is equipped with a *Scene* RGB camera and a *DepthPlanar* camera, both at 780×1240 resolution and a 70° FoV, enabling capture of RGB frames and corresponding depth GT. The pipeline is resolution-agnostic, as all components resize inputs internally or operate on metric 3D coordinates; the field of view is the primary camera-dependent factor, affecting ground coverage and depth model perspective. Mission trajectories are defined in QGroundControl [41], providing waypoint-level control over position and altitude. Coordinate consistency is ensured through identical OriginGeoPoint assignment and 90° rotation alignment between Cosys-AirSim and Unreal Engine.

c) Coordinate Frame Alignment: Establishing metric correspondence between DEM elevations and simulated depth requires careful reference frame transformation:

- **Geoid correction**: Although Tinitaly DEM is referenced to the WGS84 datum and projected in UTM Zone 32N, elevations are provided as orthometric heights (mean sea level). The EGM2008 geoid undulation model is applied to convert orthometric heights into WGS84 ellipsoidal heights, consistent with the vertical reference used by Cosys-AirSim's GNSS simulation.
- **Frame convention**: The DEM is distributed in UTM Zone 32N projected coordinates, whereas Cosys-AirSim uses a North-East-Down (NED) frame; therefore, a coordinate transformation is required.
- **Origin alignment**: The DEM origin is set as the UAV takeoff position, creating a unified local Cartesian frame where DEM queries and depth measurements share the same reference point.

This configuration ensures pixel-perfect depth, with GT georeferenced up to the accuracy of Google photogrammetry, fully preserving geographic correspondence with operational DEM databases and satisfying all requirements for DEM-based rescaling validation identified in Section I.

D. Flight Mission Configuration

Experiments followed a fixed trajectory over central Milan (45.464°N, 9.190°E) at multiple altitudes to assess altitude-dependent effects. The main evaluation was conducted at 70 m Above Ground Level (AGL) with a $-45°$ camera pitch. This configuration is representative of typical urban UAV inspections: 70 m AGL reflects a common operational altitude, while a $-45°$ oblique pitch balances ground visibility and scene context, avoiding near-nadir compression and preserving sufficient ground coverage for DEM-based rescaling. The trajectory spans areas of varying density and occlusion, producing diverse ground visibility, and was repeated at 50, 70, 100, 120, and 170 m AGL.

V. Results

This section presents the empirical findings of the DEM-rescaling pipeline. Results are organized to address each evaluation protocol defined in Section III.

A. Depth Model Performance

Figure 3 shows how the depth model performance varies with mean scene depth, which serves as a proxy for different altitude–pitch combinations on the SkyScenes dataset.

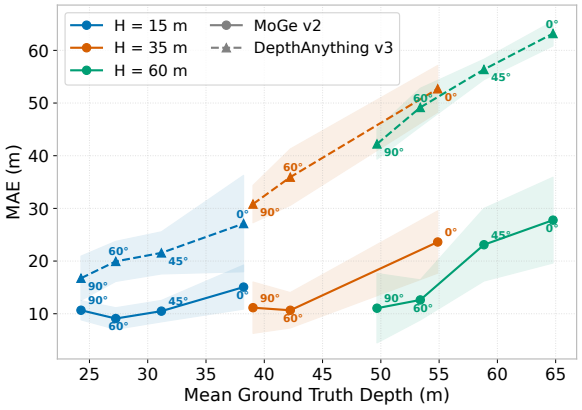


Figure 3: Depth model MAE as a function of mean scene depth across different altitudes and pitch angles on the SkyScenes dataset. Both models exhibit an approximately linear increase in error with increasing scene depth. DA v3 (dashed line) consistently produces higher errors than MoGe v2 (solid line).

TABLE I: Training data distribution for critical depth ranges and aerial scene coverage.

Category	MoGe v2	%	DA v3	%
Depth Range Distribution				
0–15 m	∼4.5M frames (Taskonomy, indoor)	50.6%	∼22K frames (scattered)	1.7%
15–50 m	∼2.1M frames	23.6%	∼8K frames	0.6%
50–100 m	∼1.9M frames (Waymo, Argoverse2)	21.3%	∼5K frames (WildRGBD)	0.4%
Aerial/Elevated Scene Coverage				
Aerial datasets	∼1.2M frames (MatrixCity, MidAir, UrbanSyn, MegaDepth)	13.5%	∼36K frames (MatrixCity, UrbanSyn, MegaDepth)	2.9%
Drone-specific	∼423K frames (MidAir)	4.7%	**None**	**0%**

Both models exhibit roughly linear MAE growth as the mean scene depth increases. The DA v3 errors rise sharply from 16 m to 65 m, while MoGe v2 increases more moderately from 10 m to 28 m over the same depth range. This gap reflects differences in training data distribution and scale modeling, as shown in Table I.

MoGe v2 is trained primarily on metrically supervised real-world data (73.6%), with substantial UAV-relevant coverage (45% in 15–100 m) and limited UAV-specific data (MidAir, 4.7%). In contrast, DA v3 relies mostly on synthetic and pseudo-labeled data (87%), with minimal aerial coverage (2.9%) and no drone-specific frames. Despite its limited aerial supervision, MoGe v2 consistently outperforms DA v3, though it still exhibits non-negligible errors.

Table II shows that at 70 m on the Synthetic-to-Geo-Real dataset, MoGe v2 produces metrically plausible estimates (mean 69.4 m vs GT 78.6 m, requiring 1.1× rescaling), whereas DA v3 suffers severe compression (mean 1.97 m, MAE 76.6 m, requiring 40× rescaling). Both networks are evaluated beyond their training depth range, and the resulting differences in depth distribution directly account for the variations in raw metric error reported in Table III.

TABLE II: Depth range statistics at 70 m altitude on Synthetic-to-Geo-Real dataset.

Depth Model	Predicted depth (m)			GT depth (m)		
	Min	Max	Mean	Min	Max	Mean
DA v3	1.17	3.66	1.97	43.29	158.38	78.59
MoGe v2	38.00	141.33	69.41			

TABLE III: Raw, normalized, and GT-Ideal rescaled MAE for DA v3 and MoGe v2 at 70 m altitude.

Model	Raw MAE (m)	Norm. MAE	GT-Ideal MAE (m)
DA v3	76.62	0.03	2.53
MoGe v2	9.40	0.03	2.30

Table III reports the raw, normalized, and GT-Ideal rescaled MAE using the methodology described in Section III-B. In raw metric terms, as confirmed by SkyScenes, MoGe v2 outperforms DA v3 (9.4 m vs 76.6 m), reflecting its superior metric depth estimation. Despite this difference, both models preserve relative depth relationships, achieving similar high relative accuracy, with a normalized MAE of ~ 0.03 m. After applying GT-Ideal based rescaling, the MAE for both models converges to 2.3–2.5 m, representing the theoretical lower bound for DEM-rescaling, constrained by intrinsic geometric distortions that cannot be corrected by a global affine transformation.

B. Ground Segmentation Quality

This section evaluates ground segmentation performance across three representative methods, highlighting their differing sensitivities to depth model characteristics and flight conditions. Table IV summarizes trajectory-averaged metrics across all configurations of the SkyScenes dataset.

EoMT consistently outperforms geometric approaches, achieving the best overall performance (IoU 0.404, Precision 0.532, Recall 0.558, F1 0.518) with low variability (IoU std 0.044), remaining stable across altitudes and pitch angles as shown in Figure 4. In contrast, geometric methods exhibit higher variance (IoU std 0.073–0.131) and stronger dependence on depth quality. CSF is the most sensitive method, favoring recall over precision: with MoGe v2, it reaches high

TABLE IV: Ground segmentation performance on SkyScenes (mean ± std). CSF and RANSAC are evaluated using the GT, DA v3, and MoGe v2 depth maps; EoMT is RGB-only.

Method	IoU	Precision	Recall	F1
CSF				
+ GT Depth	0.263 ± 0.131	0.398 ± 0.113	0.471 ± 0.274	0.374 ± 0.163
+ DA v3 Depth	0.113 ± 0.084	0.232 ± 0.178	0.261 ± 0.147	0.176 ± 0.127
+ MoGe v2 Depth	**0.294 ± 0.129**	**0.412 ± 0.141**	**0.574 ± 0.248**	**0.415 ± 0.146**
Multi-plane RANSAC				
+ GT Depth	0.284 ± 0.089	0.371 ± 0.094	0.458 ± 0.124	0.382 ± 0.103
+ DA v3 Depth	0.246 ± 0.110	0.315 ± 0.142	0.444 ± 0.137	0.326 ± 0.121
+ MoGe v2 Depth	**0.285 ± 0.073**	**0.389 ± 0.085**	**0.455 ± 0.108**	**0.389 ± 0.089**
EoMT				
RGB-based	**0.404 ± 0.044**	**0.532 ± 0.037**	**0.558 ± 0.064**	**0.518 ± 0.041**

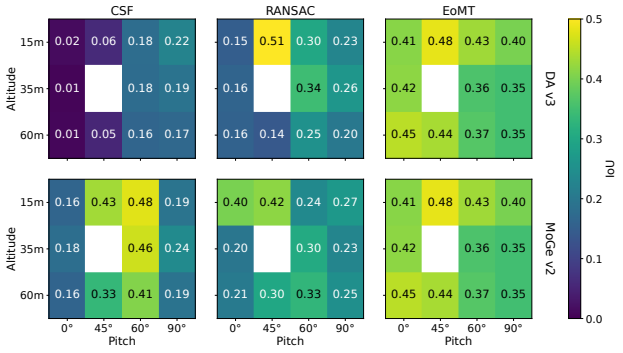


Figure 4: Segmentation IoU heatmaps across altitudes and pitch for CSF, RANSAC, and EoMT with DA v3 and MoGe v2 depths. EoMT shows consistently higher and more stable performance.

TABLE V: Ground segmentation performance on the Synthetic-to-Geo-Real dataset (mean ± std across Frames 170, 205, and 230). Best values in **bold**.

Method	IoU	Precision	Recall	F1
CSF				
+ DA v3 Depth	0.832 ± 0.054	0.835 ± 0.057	**0.997 ± 0.002**	0.908 ± 0.032
+ MoGe v2 Depth	0.912 ± 0.043	0.965 ± 0.009	0.948 ± 0.042	**0.956 ± 0.026**
Multi-plane RANSAC				
+ DA v3 Depth	0.891 ± 0.040	0.955 ± 0.008	0.929 ± 0.036	0.941 ± 0.022
+ MoGe v2 Depth	0.889 ± 0.079	0.972 ± 0.008	0.918 ± 0.088	0.942 ± 0.049
EoMT				
RGB-based	**0.906 ± 0.037**	**0.974 ± 0.014**	0.932 ± 0.053	0.934 ± 0.053

ground coverage (Recall 0.574) at the expense of false positives, while DA v3 significantly degrades performance (IoU 0.113). RANSAC shows more balanced robustness, with a narrower IoU range (0.246–0.285) across depth sources. Results on the Synthetic-to-Geo-Real dataset in Table V confirm these patterns in realistic georeferenced scenarios.

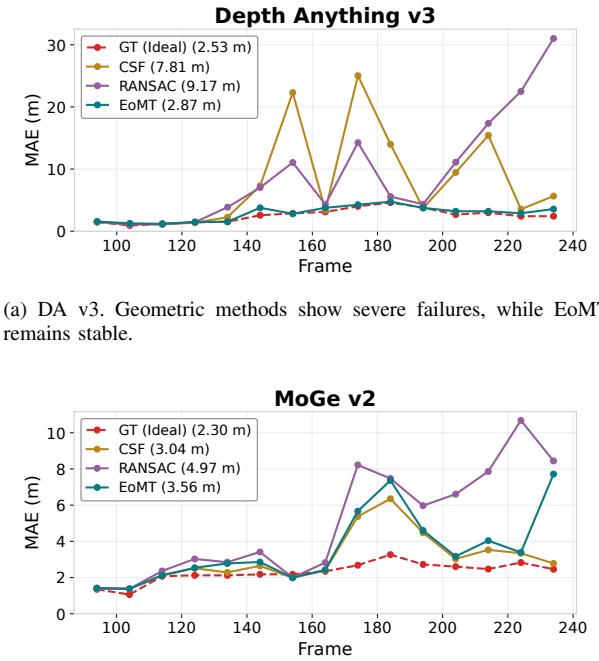
C. DEM-Rescaling: Depth & Segmentation Effects

Table VI reports the mean trajectory-level depth estimation across the complete test sequence. Results show that both the segmentation method and depth model quality strongly influence the rescaled depth quality. With the realistic depth range of MoGe v2, both EoMT and CSF achieve high accuracy (MAE < 3.6 m, $\delta_{1.25} > 99.6\%$), while RANSAC performs worse (MAE $\sim$7 m, $\delta_{1.25} \sim 90\%$) due to its sensitivity to fixed geometric thresholds along the trajectory. Under the compressed range of DA v3, performance diverges: EoMT maintains robust accuracy, while CSF and RANSAC degrade significantly (7.17 m and 10.93 m MAE, $\delta_{1.25}$ dropping to 83.90% and 75.82%). To better understand the variability behind these aggregate metrics, Figure 5 reports the per-frame MAE along the evaluated sequence, highlighting critical differences in how segmentation methods respond to scene variations along the flight trajectory.

CSF with DA v3 exhibits high-impact errors, with 22–25 m spikes between frames 150–180. With MoGe v2, CSF stabilizes at a 3.04 m mean but still reaches 6.3 m peaks. RANSAC shows consistent frame-to-frame instability, with

TABLE VI: Mean trajectory-level performance for different segmentation methods using DA v3 and MoGe v2 depth inputs.

Segmentation	MAE (m) ↓		ABS rel % ↓		$\delta_{1.25}$ % ↑	
	DA v3	MoGe v2	DA v3	MoGe v2	DA v3	MoGe v2
CSF	7.17	**3.05**	10.20	**4.20**	83.90	**99.78**
RANSAC	10.93	6.95	16.56	10.24	75.82	89.73
EoMT	**2.87**	3.57	**3.94**	5.03	**99.43**	99.66



(a) DA v3. Geometric methods show severe failures, while EoMT remains stable.

(b) MoGe v2. Improved depth calibration stabilizes CSF, but RANSAC remains unstable.

Figure 5: Per-frame MAE along the 70 m trajectory.

peak errors of 31.1 m for DA v3 and 10.6 m for MoGe v2. EoMT remains remarkably stable across the trajectory for both depth models.

Figure 6 evaluates rescaling robustness across the operational altitude range of 50–170 m, within which the observable depth range naturally expands.

EoMT consistently maintains stable and accurate performance with both depth models, achieving MAE values of roughly 3–6 m, closely matching the ideal rescaling achievable with GT depth data, across all altitudes and depth variations. In contrast, CSF performance degrades sharply with DA v3 due to depth compression, with MAE increasing by about 5× from low to high altitudes, whereas it remains stable with MoGe v2 (MAE $\sim$3.6–4.4 m). RANSAC shows intermediate but less predictable performance, with MAE values ranging from 7 to 16 m and irregular fluctuations.

D. Spatial Ground Sampling Strategy

This section analyzes how ground-pixel distribution and coverage affect rescaling accuracy.

Table VII shows that at low altitudes (50–70 m), central subregions (r2–r3, c1–c2) consistently achieve the lowest

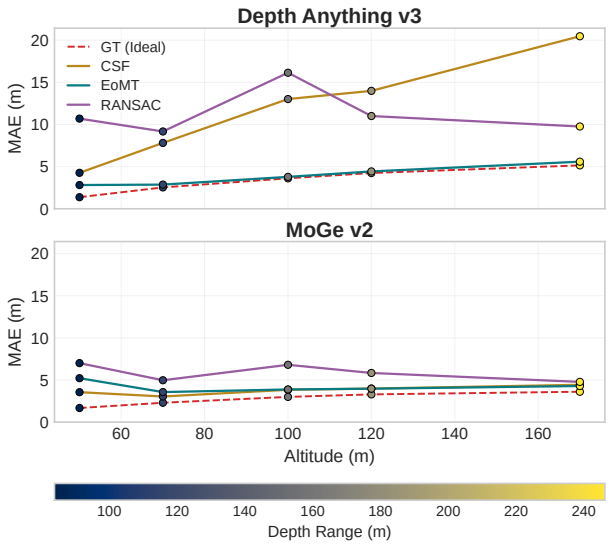


Figure 6: Rescaling MAE over the 50–170 m altitude range for DA v3 (top) and MoGe v2 (bottom).

TABLE VII: MAE (m) for 4×4 subregions at various altitudes using EoMT + DA v3. Average land cover (%) is shown below each row-column identification. Green cells indicate subregions outperforming full ground; red cells indicate MAE >2× baseline. Dashes (–) indicate subregions with no ground pixels within the fixed grid partition.

Altitude (m)	r0c1	r1c0	r1c1	r1c2	r2c0	r2c1	r2c2	r3c0	r3c1	r3c2	full
	4.1%	4.4%	8.2%	7.1%	8.5%	9.0%	15.6%	9.8%	15.1%	14.9%	100%
50	2.37	2.26	3.14	3.32	1.56	2.02	1.57	3.00	2.20	1.59	1.91
70	6.68	4.17	4.03	3.40	3.42	3.37	4.06	3.48	3.27	3.61	3.77
100	3.85	14.18	4.36	13.18	–	3.78	3.66	–	4.11	4.19	3.33
120	9.84	25.23	7.32	–	29.04	4.15	4.25	–	4.51	5.45	4.31
170	9.99	–	6.28	–	6.75	5.84	5.72	–	6.23	6.47	5.48

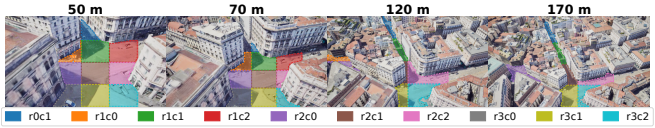


Figure 7: Visualization of the 4×4 subregion grid at different altitudes.

MAE, outperforming the full-ground baseline by up to 18% at 50 m and 13% at 70 m. At higher altitudes, performance becomes more variable as ground visibility decreases, though central regions remain robust within 10% of the baseline. At 120 m, two central subregions outperform the full-ground baseline by 2.5% despite covering only a small fraction of ground pixels.

E. Computational Efficiency

Table VIII reports the mean segmentation latency of each method on a ground station with an NVIDIA RTX A3000 GPU. Segmentation accounts for 66–69% of the total pipeline time for CSF. EoMT achieves a 6.5× speedup compared to CSF, even though it runs on the CPU (93 ms vs 603–675 ms). RANSAC provides the highest efficiency (60–70 ms) but with accuracy degradation.

TABLE VIII: Complete pipeline timing with alternative segmentation methods and depth models.

Depth Model	Segmentation Method	Depth (ms)	Rescaling (ms)	Segmentation (ms)	Total (ms)	FPS	Seg % Pipeline
DA v3	CSF	308.00	1.83	603.12	912.95	1.09	66.1%
	RANSAC			59.98	369.81	2.70	16.2%
	EoMT			93.18	403.01	2.48	23.1%
MoGe v2	CSF	294.59	2.57	675.01	972.17	1.03	69.4%
	RANSAC			68.64	365.80	2.73	18.8%
	EoMT			93.18	390.34	2.56	23.9%

VI. Discussion

Section V analyzed how depth model characteristics and ground segmentation strategies affect DEM-based metric scale recovery in UAV scenarios. The ideal rescaling analysis (Table III) shows that both MoGe v2 and DA v3 achieve comparable intrinsic accuracy under perfect affine correction, despite markedly different raw depth scales. This suggests that the main challenge is not only depth model selection, but also reliably estimating transformation parameters through correct ground segmentation, highlighting the central role of ground quality in metric scale recovery. Building on this, CSF used in TanDepth, along with other geometric methods operating on 3D point clouds, results in a strong coupling between depth and segmentation. This coupling degrades under compression or calibration artifacts, precisely when rescaling is most critical, leading to scene-dependent instability and scenario-specific parameter tuning. In contrast, semantic segmentation leverages depth-independent visual cues, providing stable metric performance across depth models and enabling robust operation in conditions where geometric methods fail. Further investigation of subregions reinforces the importance of segmentation quality over spatial extent. Compact, high-precision regions that cover only 7–16% of the ground consistently outperform full-ground baselines by avoiding common failure sources such as boundary depth degradation, occlusions, and facade misclassification. This shows that selective sampling can achieve favorable accuracy–efficiency trade-offs under limited ground visibility, although the optimal subregion depends on altitude and spatial location.

VII. Conclusion and Future Directions

This work lays the foundation for robust DEM-based rescaling in operational UAV pipelines by enforcing metric consistency in monocular depth estimation. The proposed Synthetic-to-Geo-Real framework bridges the gap between controlled synthetic validation and real-world georeferenced deployment, providing a controlled and reproducible evaluation protocol for UAV perception. Our systematic analysis reveals that the critical limitation in DEM-based rescaling lies in the robustness of the grounding strategy used to recover metric scale. By integrating semantic segmentation into the rescaling pipeline, we achieve up to a 60% reduction in MAE and speed up the computation compared to previously proposed state-of-the-art geometric filters, such

as CSF. These results position EoMT as the most promising candidate for deployment on embedded UAV platforms. The proposed design enables stable and model-agnostic metric recovery, even in the presence of the intrinsic inaccuracies of modern MDE models. Strategic ground sampling focused on compact, high-confidence subregions matches or improves full-ground performance while reducing computational cost. Future work will extend this pipeline from per-frame estimation to temporal grounding, enabling continuous, trajectory-aware metric perception even in zero-ground scenarios.

ACKNOWLEDGMENTS

This work was supported by a PhD fellowship from Leonardo S.p.A. Rendering via Google 3D Tiles API.

REFERENCES

[1] R. Sapkota, K. I. Roumeliotis, and M. Karkee, "Uavs meet agentic ai: A multidomain survey of autonomous aerial intelligence and agentic uavs," 2025.

[2] U. Rajapaksha, F. Sohel, H. Laga, D. Diepeveen, and M. Bennamoun, "Deep learning-based depth estimation methods from monocular image and videos: A comprehensive survey," *ACM Computing Surveys*, vol. 56, p. 1–51, Oct. 2024.

[3] B. Wu *et al.*, "Vgld: Visually-guided linguistic disambiguation for monocular depth scale recovery," *arXiv preprint arXiv:2505.02704*, 2025. Language-based scale recovery using textual descriptions and object-scene associations.

[4] Xu *et al.*, "Monocular absolute depth estimation from motion for small unmanned aerial vehicles by geometry-based scale recovery," *Sensors*, vol. 24, no. 14, p. 4541, 2024. Geometry-based scale recovery using navigation sensor pose data and multi-view geometry.

[5] R. Giubilato, S. Chiodini, M. Pertile, and S. Debei, "Scale correct monocular visual odometry using a lidar altimeter," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3694–3700, 2018.

[6] H. Florea and S. Nedevschi, "Tandepth: Leveraging global dems for metric monocular depth estimation in uavs," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, p. 5445–5459, 2025.

[7] W. Zhang, J. Qi, P. Wan, H. Wang, D. Xie, X. Wang, and G. Yan, "An easy-to-use airborne lidar data filtering method based on cloth simulation," *Remote Sensing*, vol. 8, no. 6, 2016.

[8] Y. Lyu, G. Vosselman, G. Xia, A. Yilmaz, and M. Y. Yang, "Uavid: A semantic segmentation dataset for uav imagery," 2020.

[9] H. Florea, V.-C. Miclea, and S. Nedevschi, "Wilduav: Monocular uav dataset for depth estimation tasks," *2021 IEEE 17th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2021.

[10] F. Nex, E. K. Stathopoulou, and F. Remondino, "UseGeo - A UAV-based multi-sensor dataset for geospatial research," *International Journal of Applied Earth Observation and Geoinformation*, vol. 132, p. 103851, 2024.

[11] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, "Tartanair: A dataset to push the limits of visual slam," 2020.

[12] G. Rizzoli, F. Barbato, M. Caligiuri, and P. Zanuttigh, "Syndrone–multi-modal uav dataset for urban scenarios," *arXiv preprint arXiv:2308.10491*, 2023.

[13] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," 2017.

[14] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," 2014.

[15] J. Lin *et al.*, "Depth3: Recovering metric depth from a single image," *arXiv*, 2025.

[16] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," 2019.

[17] H. Hu, Y. Feng, D. Li, S. Zhang, and H. Zhao, "Monocular depth estimation via self-supervised self-distillation," *Sensors*, vol. 24, no. 13, 2024.

[18] R. Birkl, D. Wofk, and M. Müller, "Midas v3.1 – a model zoo for robust monocular relative depth estimation," 2023.

[19] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," 2023.

[20] R. Wang, S. Xu, Y. Dong, Y. Deng, J. Xiang, Z. Lv, G. Sun, X. Tong, and J. Yang, "Moge-2: Accurate monocular geometry with metric scale and sharp details," 2025.

[21] H. Lin, S. Chen, J. H. Liew, D. Y. Chen, Z. Li, G. Shi, J. Feng, and B. Kang, "Depth anything 3: Recovering the visual space from any views," *arXiv preprint arXiv:2511.10647*, 2025.

[22] H. Florea and S. Nedevschi, "Survey on monocular depth estimation for unmanned aerial vehicles using deep learning," in *2022 IEEE 18th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 319–326, 2022.

[23] M. Irfan, S. Dalai, P. Trslic, J. Riordan, and G. Dooly, "Lsaf-lstm-based self-adaptive multi-sensor fusion for robust uav state estimation in challenging environments," *Machines*, vol. 13, no. 2, 2025.

[24] C. Zhang, X. Weng, Y. Cao, and M. Ding, "Monocular absolute depth estimation from motion for small unmanned aerial vehicles by geometry-based scale recovery," *Sensors*, vol. 24, no. 14, 2024.

[25] Y. Zhou, X. Li, S. Li, X. Wang, and Z. Shen, "Ground-vio: Monocular visual-inertial odometry with online calibration of camera-ground geometric parameters," 2023.

[26] R. Zhu, C. Wang, Z. Song, L. Liu, T. Zhang, and Y. Zhang, "Scaledepth: Decomposing metric depth estimation into scale prediction and relative depth estimation," 2024.

[27] T. Gomes, D. Matias, A. Campos, L. Cunha, and R. Roriz, "A survey on ground segmentation methods for automotive lidar sensors," *Sensors*, vol. 23, no. 2, 2023.

[28] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, pp. 381–395, 1981.

[29] T. Rabbani, F. Heuvel, and G. Vosselman, "Segmentation of point clouds using smoothness constraint," *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 36, 01 2006.

[30] K. Zhang, S.-C. Chen, D. Whitman, M.-L. Shyu, J. Yan, and C. Zhang, "A progressive morphological filter for removing nonground measurements from airborne lidar data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 4, pp. 872–882, 2003.

[31] T. Kerssies, N. Cavagnero, A. Hermans, N. Norouzi, G. Averta, B. Leibe, G. Dubbelman, and D. de Geus, "Your ViT is Secretly an Image Segmentation Model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[32] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," 2016.

[33] S. Khose, A. Pal, A. Agarwal, Deepanshi, J. Hoffman, and P. Chattopadhyay, "Skyscenes: A synthetic dataset for aerial scene understanding," 2024.

[34] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 1–16, 2017.

[35] Cesium GS, "Cesium for unreal." `https://github.com/CesiumGS/cesium-unreal`.

[36] "Photorealistic 3d tiles — learn — cesium." `https://cesium.com/learn/photorealistic-3d-tiles-learn/`.

[37] Istituto Nazionale di Geofisica e Vulcanologia (INGV), "Tinitaly/01 — italian national digital elevation model." `https://tinitaly.pi.ingv.it/`.

[38] S. Tarquini and L. Nannipieri, "The 10m-resolution tinitaly dem as a trans-disciplinary basis for the analysis of the italian territory: Current trends and new perspectives," *Geomorphology*, vol. 281, pp. 108–115, 2017.

[39] W. Jansen, E. Verreycken, A. Schenck, J.-E. Blanquart, C. Verhulst, N. Huebel, and J. Steckel, "Cosys-airsim: A real-time simulation framework expanded for complex industrial applications," in *2023 Annual Modeling and Simulation Conference (ANNSIM)*, pp. 37–48, 2023.

[40] L. Meier, D. Honegger, and M. Pollefeys, "Px4: A node-based multithreaded open source robotics framework for deeply embedded platforms," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6235–6240, 2015.

[41] QGroundControl Development Team, "Qgroundcontrol." `https://github.com/mavlink/qgroundcontrol`.