Monocular Depth Estimation from UAV Images for 3D Documentation of Architectural Heritage: A Depth Anything V2-Based Approach

Filiberto Chiabrando[1], Francesca Gallitto[2], Andrea Maria Lingua[2], Stefania Manca[2], Alessio Martino[1], Francesca Matrone[2], Alessandra Spadaro[2]

[1] DAD, Politecnico di Torino, Viale Pier Andrea Mattioli 39, 10125 Torino, Italy - (filiberto.chiabrando, alessio.martino)@polito.it
[2] DIATI, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy - (francesca.gallitto, andrea.lingua, stefania.manca, francesca.matrone, alessandra.spadaro)@polito.it

Keywords: Monocular depth estimation, UAV, architectural heritage, fine-tuning, photogrammetry, domain adaptation

Abstract

Monocular depth estimation (MDE) has reached notable maturity in computer vision, yet its application to UAV-based architectural heritage documentation remains underexplored. This study assesses whether the depth foundation model Depth Anything V2 can be transferred from terrestrial to aerial imagery. The analysis relies on MDE4BH, a benchmark of over 3,000 UAV images covering ten heterogeneous heritage scenarios (urban areas, façades, towers, villas, domes, and archaeological sites). Masked photogrammetric depth maps serve as metric reference for calibration, validation, and supervised retraining. Two baseline configurations are evaluated: a relative model with scene-specific linear rescaling and the direct application of the metric model. The rescaled relative model shows acceptable performance in several subsets, whereas the metric model exhibits systematic bias, weak consistency, and scale collapse due to domain shift between terrestrial training data and aerial acquisition geometry. To address these limitations, a two-step fine-tuning strategy is introduced, focusing on the decoder and regression head. The first stage uses mainly oblique UAV images; the second integrates oblique and nadir views to improve viewpoint generalization. The adapted model significantly reduces bias and enhances metric stability across the benchmark. However, residual errors remain spatially structured, with clustering and recurrent artefacts near object boundaries, multi-level roofs, and radiometrically heterogeneous surfaces. Although accuracy is still insufficient for demanding metric applications, the results support the use of MDE as a complementary source for thematic interpretation, scene understanding, robotics, navigation, and related tasks where strict geometric precision is not required.

1. Introduction

Monocular depth estimation (MDE) has been studied for nearly two decades and has recently reached a significant level of maturity, driven by advances in deep learning, transformer-based dense prediction, and large-scale vision foundation models. MDE aims to infer a dense depth map from a single RGB image, without relying on stereo geometry or multi-view triangulation. In contrast to photogrammetric depth, which is derived from image correspondences and camera geometry, monocular depth is inferred from learned visual cues such as perspective, texture, shading, and semantic context. Early learning-based methods demonstrated the feasibility of the problem, while later deep architectures substantially improved accuracy and generalization (Saxena et al., 2005; Eigen et al., 2014; Laina et al., 2016; Ranftl et al., 2021; Ranftl et al., 2022; Yang et al., 2024a; Yang et al., 2024b, Ke et al., 2024, Bochkovskii et al., 2024, Wang et al, 2025).

These developments are of particular interest for architectural heritage documentation, where image-based workflows increasingly benefit from dense geometric information. In this context, MDE is not intended to replace photogrammetric reconstruction, but to be evaluated as a complementary source of geometric information for tasks such as 3D reconstruction support, orthoimage production, semantic interpretation, HBIM-oriented modelling, and GIS integration. Its practical value, however, depends not only on visual plausibility, but also on metric coherence and robustness across different acquisition conditions.

This issue becomes critical for UAV imagery. Most recent MDE models have been trained mainly on terrestrial datasets, with near-ground viewpoints and scene compositions that differ substantially from those of drone-based heritage surveys. UAV acquisitions typically combine nadir and oblique views, strong variations in relative flight height, abrupt transitions between roofs, façades, sky, vegetation, and terrain, and geometrically complex architectural forms. Under these conditions, the priors learned from terrestrial imagery may no longer be fully valid, leading to both local artefacts and systematic scale distortions in metric prediction.

A further limitation concerns benchmark data. Commonly used datasets such as KITTI, NYU Depth v2, ETH3D, and DIML do not adequately represent the geometric and radiometric conditions of UAV-based heritage documentation. More recent resources, including ArchDepth and UseGeo, are closer to this application domain, but still only partially cover the combination addressed here, namely heterogeneous heritage objects observed under mixed nadir and oblique UAV configurations and evaluated against dense photogrammetric depth maps suitable for metric assessment and supervised adaptation (Geiger et al., 2013; Silberman et al., 2012; Schöps et al., 2017; Cho et al., 2021; Stathopoulou and Remondino, 2022; Nex et al., 2024).

Within this framework, the present study investigates whether a recent foundation model for monocular depth estimation can be adapted from terrestrial vision to UAV-based architectural heritage documentation. The analysis focuses on Depth Anything V2, selected because it is among the most competitive recent MDE frameworks, provides both relative and metric variants within a unified architecture, and offers a suitable basis for analysing both depth ordering and metric consistency in the UAV domain. Using a benchmark of more than 3,000 drone images acquired over ten heterogeneous heritage case studies, photogrammetric depth maps generated in a commercial Structure from Motion (SfM) software are employed as masked

metric ground truth for calibration, validation, and supervised retraining. The paper analyses the behaviour of the baseline relative and metric models on aerial imagery, introduces a two-step fine-tuning strategy focused on the upper decoder and depth regression head, and discusses the potential of MDE as a complementary tool for heritage-oriented 3D documentation workflows.

2. 2. Related work

As outlined above, the literature most relevant to this study can be grouped into three areas: robust relative depth estimation, metric depth estimation, and benchmark data.

A first area concerns relative monocular depth estimation. Major improvements have been achieved through heterogeneous training data and stronger dense prediction architectures. MiDaS showed that mixing datasets can improve zero-shot transfer, while DPT highlighted the effectiveness of transformer-based architectures for dense prediction. More recently, Depth Anything and Depth Anything V2 further strengthened robustness and generalization through large-scale data curation and training, establishing strong baselines for generic monocular depth prediction (Ranftl et al., 2022; Ranftl et al., 2021; Yang et al., 2024a; Yang et al., 2024b).

A second area concerns metric depth estimation, which remains more challenging because it requires scale-consistent predictions across different scenes and acquisition settings. ZoeDepth and UniDepth represent important advances in this direction. However, their performance still depends strongly on domain compatibility, which is particularly problematic for UAV imagery characterized by top-down viewpoints, large depth variations, and complex architectural geometries (Bhat et al., 2023; Piccinelli et al., 2024).

A third area concerns the datasets used to train and evaluate MDE models. Widely adopted benchmarks such as KITTI, NYU Depth v2, ETH3D, and DIML have been fundamental to the field, but they mainly represent automotive, indoor RGB-D, or generic multi-view settings. As a result, they do not adequately describe UAV-based heritage documentation. Closer references are ArchDepth, introduced in a photogrammetric context, and UseGeo, a UAV-based multi-sensor dataset for geospatial research. Nevertheless, these resources still only partially reflect the heterogeneous heritage scenes and mixed aerial viewpoints considered in the present work (Geiger et al., 2013; Silberman et al., 2012; Schöps et al., 2017; Cho et al., 2021; Stathopoulou and Remondino, 2022; Nex et al., 2024).

Against this background, the present work focuses on the adaptation of Depth Anything V2 to UAV-based architectural heritage imagery, with particular attention to the metric limitations that arise when models developed primarily for terrestrial vision are transferred to aerial documentation scenarios (Yang et al., 2024b).

3. MDE4BH Benchmark

To address the limitations outlined above, a dedicated benchmark, termed MDE4BH (Monocular Depth Estimation for Built Heritage), was defined for the evaluation of monocular depth estimation in UAV-based architectural heritage documentation. It was designed to capture the diversity of viewpoints, object morphologies, and survey conditions that characterise real heritage acquisitions and that remain underrepresented in existing benchmarks.

ID	Image	Drone focal length pixel size	n. images	n.im/H_{rel} [m]
(A) Timespan Museum (Scozia)		DJI Matrice M4E 12.29 mm 3.36 μm	n. 80 nadir. n. 130 obl. 60 ° n. 290 obl. 45°	150/20 350/40
(B) Venaria roof		DJI Matrice 350 35 mm 4.4 μm	n. 250 nadir. n. 180 obl. 45°	250/60 180/90
(C) Venaria facade		DJI Mini 3 Pro 6.72 mm 2.4 μm	n.50 obl. 90° n.25 obl. 60°	18 m
(D) Chianca Tower		DJI Mavic 3M 4.34 mm 3.3 μm	n. 40 obl. 45° n. 55 obl. 65°	40/20 55/5
(E) Sacro Monte di Crea		DJI Mini 3 Pro 6.72 mm 2.4 μm	n. 90 nadir. n. 95 obl. 90° n. 30 obl. 45° n. 115 obl. 60°-80°	330/5
(F) Villino Caprifoglio		DJI Spark 4.4 mm 1.55 μm	n. 70 nadir. n.105 pitch 30°-70° n.25 pitch 90°	130/25 70/12
(G) Mappano (TO)		DJI Phantom 4 Pro 8.8 mm 2.41 μm	n. 195 nadir. n. 220 obl. 45°	195/80 220/60
(H) Dome of San Lorenzo		DJI Mini 3 Pro 6.72 mm 2.4 μm	n. 10 nadir. n.30 obl. 40°-60° n.90 obl. 60°-90°	130/13
(I) Valentino Castle (TO)		DJI Phantom 4 Pro 8.8 mm 2.41 μm	n. 209 nadir. n. 67 obl. 80° n. 205 obl. 45°	209/40 272/60
(L) Bagno Arabo di Mezzagno ne (RG)		DJIMini mavic 3 4.49 mm 1.76 μm	n. 229 nadir. n. 139 obl. 45°	368/18

Table 1 - Composition of the MDE4BH benchmark used in the experiments

The current release includes more than 3,000 UAV images collected over ten case studies, covering urban blocks, monumental façades, isolated towers, villas, a dome, an extensive architectural complex, and archaeological remains. The datasets include nadir, oblique, and mixed-view imagery, with relative flight heights ranging from about 5 m to 90 m and different DJI acquisition platforms. This heterogeneity was intentionally preserved in order to analyse model behaviour under varying geometric and radiometric conditions rather than within a single narrowly controlled scenario (Table 1).

Metric reference depth maps were generated in a commercial SfM software using full-resolution processing and Ultra High dense reconstruction. The depth maps were subsequently masked to remove sky, voids, reflective surfaces, and other unreliable areas. This enabled the use of photogrammetric depth as common

metric ground truth for calibration, validation, residual analysis, and supervised adaptation of the tested models.

In this sense, MDE4BH was conceived not simply as a collection of UAV images, but as a benchmark specifically structured to evaluate the transfer of monocular depth estimation models from terrestrial vision to heritage-oriented aerial documentation.

4. Experimental Protocol and Baseline Analysis

4.1 Baseline application of Depth Anything V2 models

The experimental workflow first evaluates two baseline conditions. In the first, the relative Depth Anything V2 base model is applied and then linearly rescaled against the masked photogrammetric ground truth. In the second, the metric outdoor Depth Anything V2 model is applied directly to UAV imagery. This two-baseline design is important because it separates the problem of recovering a meaningful depth ordering from the problem of transferring metric scale to aerial heritage scenes.

For the relative configuration, the conversion from the predicted relative depth (or disparity-like map) to metric depth follows a masked linear regression (with scale a and shift b) against the ground truth over valid pixels only estimated using least square principle:

$$\widehat{d_i} = a\widehat{dr_i} + b$$

$$\left\|\widehat{d_i} - d_i\right\|^2 = \left\|a\widehat{dr_i} + b - d_i\right\|^2 = min \quad (1)$$

where: $\widehat{dr_i}$ are the relative depths, d_i are the real depths of ground truth, $\widehat{d_i}$ are the predicted depths (in this case rescaled and shifted). This allow to examine how much of the domain gap can be absorbed by scale and offset recovery alone.

Evaluation relies on residuals mean (m), residuals median (M), residuals standard deviation (std), the logarithmic RMSE ($RMSE_{log}$) and Absolute Relative Error (AbsRel).

AbsRel measures the mean absolute prediction error normalized by the reference depth,

$$AbsRel = \frac{1}{n}\sum_{i=1}^{n}\frac{|d_i - \widehat{d_i}|}{d_i} \quad (3)$$

useful to analyse when scenes span a broad depth range, since it expresses the error in relative rather than absolute terms.

RMSE log is defined as

$$RMSE_{log} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log d_i - \log \widehat{d_i})^2} \quad (4)$$

and evaluates the discrepancy in logarithmic space, thus being more sensitive to relative scale inconsistencies than to absolute metric offsets. The combined use of AbsRel and RMSE log is particularly useful because it allows one to assess both the overall relative depth accuracy and the tendency of the model to produce scale compression or expansion when transferred from terrestrial training domains to UAV heritage imagery. The use of standard depth-estimation metrics follows established evaluation practice in monocular depth prediction. The results are shown in Table 2.

The properly rescaled relative model shows the most consistent behaviour in nadir imagery, where the lowest error values are observed. This is particularly evident in case G, which provides the best overall performance, with Std = 0.85 m, RMSElog = 0.02, and Abs Rel = 0.011 in the nadir subset. A similar trend is found in case A, where nadir images also produce markedly better results than oblique views, while the 60° oblique subset shows a clear degradation, reaching Std = 12.94 m, RMSElog = 1.23, and Abs Rel = 0.223.

Case	Subset	Std [m]	RMSE log	Abs Rel	Comments
A	All	6.1	0.38	0.10	Moderate overall performance
A	Nad.	1.3	0.04	0.01	Best subset
A	Obl. 45°	4.6	0.12	0.07	Acceptable degradation
A	*Obl. 60°*	*12.9*	*1.23*	*0.22*	*Critical subset*
D	All	8.6	0.15	0.10	Weak overall behaviour
D	Obl. 45°	6.0	0.08	0.06	Best in case D
D	Obl. 60°	12.6	0.19	0.13	Marked degradation
C	All	2.7	0.15	0.07	Stable overall behaviour
C	Nad.	2.1	0.11	0.05	More reliable subset
C	Obl. 60°	3.9	0.22	0.09	Moderate degradation
G	All	4.3	0.07	0.05	Best overall case
G	*Nad.*	*0.8*	*0.02*	*0.01*	*Best result*
G	Obl. 45°	6.9	0.07	0.05	Good relative stability

Table 2 - Representative baseline results before fine-tuning, using relative-model rescaling and metric-model scale collapse.

Case	Subset	m [m]	Std [m]	RMSE log	Abs Rel	Comments
A	All	-25.0	6.4	0.99	0.59	Weak overall accuracy
A	Nad.	-19.7	2.8	0.96	0.60	Limited nadir gain
A	Obl. 45°	-31.0	5.3	1.13	0.65	High relative error
A	Obl. 60°	-16.8	11.7	0.74	0.49	Large dispersion
D	All	-25.7	17.3	0.50	0.69	Very poor overall behaviour
D	Obl. 45°	-27.0	9.9	0.52	0.71	Unstable metric response
D	*Obl. 60°*	*-23.7*	*28.5*	*0.66*	*0.47*	*Extreme dispersion*
C	All	-7.1	1.5	0.50	0.37	Best overall case
C	*Nad.*	*-5.0*	*1.3*	*0.37*	*0.29*	*Best subset*
C	Obl. 60°	-11.6	2.0	0.77	0.52	Clear oblique degradation
G	*All*	*-39.4*	*3.9*	*1.49*	*0.76*	*Severe scale inconsistency*
G	Nad.	-31.8	2.8	1.17	0.67	Weak nadir behaviour
G	Obl. 45°	-45.1	4.7	1.73	0.81	Worst relative error

Table 3 - Representative baseline results before fine-tuning, using Dav2 metric model

Case C is more stable across the available subsets, with comparatively limited variation between the nadir and oblique configurations, although the oblique images still yield higher errors than the nadir ones. Case D, available only for oblique subsets in this summary, confirms the increasing difficulty associated with steeper viewing angles, with the 60° configuration performing worse than the 45° one. Overall, the results indicate that the rescaled relative model can provide satisfactory behaviour in favourable conditions, especially for nadir or moderately oblique views, whereas performance degrades as viewpoint inclination and scene complexity increase.

The direct application of the metric variant of Depth Anything V2 to UAV imagery yields markedly weaker results than the properly rescaled relative model. In general (Table 3), AbsRel values remain high across all representative cases, while RMSE

log highlights significant scale inconsistencies, confirming that the model does not preserve a reliable metric response when transferred directly from terrestrial training domains to aerial heritage scenes.

Among the analysed examples, case C is the most favourable, with the lowest overall errors and the best nadir performance (Std = 1.3 m, RMSE log = 0.37, Abs Rel = 0.29). Even in this case, however, the oblique subset shows a clear degradation. Case A presents consistently weak behaviour across all subsets, with only limited improvement in nadir views and high relative errors in all configurations. Case D is particularly critical, especially for the 60° oblique subset, where the dispersion becomes extreme (Std = 28.5 m), indicating a severe loss of metric stability. Case G further confirms this pattern, with the highest RMSElog and AbsRel values of the table, suggesting strong scale distortion even where the standard deviation is not the largest. The non-zero mean residuals reported in Table 3 further indicate that the model is affected by significant bias, not only by increased dispersion. This suggests a systematic metric offset in the predictions, consistent with the transfer difficulties observed when the metric model is directly applied to UAV imagery.

Overall, these results support the interpretation that the metric DA-V2 model suffers from a strong domain shift when applied directly to UAV heritage imagery. The problem is not limited to local errors, but affects the global metric consistency of the prediction, with frequent scale compression or misalignment across different viewpoints and scene configurations.

These representative cases are reported here for synthesis, while the same analysis was carried out over the full benchmark. The baseline results indicate that the observed limitations cannot be explained by calibration issues alone, but are also related to the scene priors learned during training. Models developed mainly on terrestrial datasets such as KITTI, NYU Depth v2, ETH3D, and DIML are exposed to viewing conditions typically structured around near-ground perspectives, moderate camera heights, and scene layouts organised with respect to the ground plane and the horizon. UAV heritage imagery departs substantially from these assumptions, owing to the presence of nadir and oblique views, stronger height variations, and more complex architectural depth configurations. In the representative cases analysed here, this domain mismatch is reflected in the weak metric consistency of the direct metric model and in the markedly better behaviour of the relative model after scene-specific rescaling. These findings suggest that, in the UAV domain, baseline metric prediction remains strongly affected by transfer limitations even when local depth ordering is only partially preserved.

4.2 Two-Step Fine-Tuning

To reduce the domain gap highlighted by the baseline analysis, a two-step supervised fine-tuning strategy was adopted for Depth Anything V2. The adaptation was designed to act primarily on the decoder and on the depth regression head, while preserving the general visual priors learned by the backbone and thus limiting overfitting (Figure 1). This choice reflects the objective of the study, namely to verify whether a recent general-purpose depth foundation model can be effectively specialised to UAV-based heritage imagery through a controlled retraining of its most task-specific components, rather than by training a new model from scratch.

The first fine-tuning stage focused exclusively on oblique UAV images, which represent both one of the most informative configurations for architectural geometry and one of the most problematic conditions for the baseline model. Approximately 1,000 images were used in this stage, with 80% for training and 20% for validation. Starting from the corrected model obtained after the first step, the second stage introduced a mixed dataset composed of about 70% oblique and 30% nadir images, again for a total of about 1,000 images, in order to improve viewpoint generalisation and reduce the imbalance between façade-oriented and top-down scene configurations.

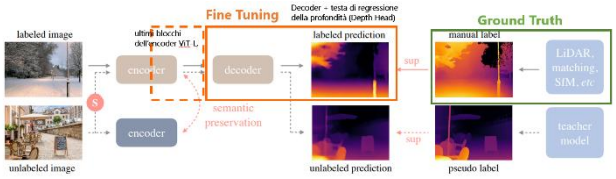

Figure 1. Components of the *Depth Anything v2* architecture involved in the fine-tuning process.

Ground-truth supervision was provided by the masked photogrammetric depth maps described above. In addition, the experimental protocol included a held-out test subset of about 10% of the dataset, composed of images not used during training, while the last two case studies were entirely excluded from the training phase. The optimisation combined an L_{L1} depth loss, a scale-invariant logarithmic loss, and a smoothness regularisation term :

$$L = \lambda_1 L_{L_1} + \lambda_2 L_{silog} + \lambda_3 L_{smooth} \tag{5}$$

weighted respectively by $\lambda_1 = 1.0$, $\lambda_2 = 0.1$, $\lambda_3 = 0.003$.

The L1 depth loss was computed as:

$$L_{L_1} = \frac{1}{n}\sum_{i=1}^{n}\left|d_i - \hat{d}_i\right| \tag{6}$$

The SIlog loss (Scale Invariant logaritmic loss) can be estimated using:

$$L_{si-log} = \frac{1}{n}\sum_{i=1}^{n}(\log d_i - \log \hat{d}_i)^2 + -\frac{\lambda_{bil}}{n^2}\left(\sum_{i=1}^{n}(\log d_i - \log \hat{d}_i)^2\right)^2 \tag{7}$$

where the second term reduces the sensitivity to global scale errors, thus making the loss more scale-invariant, with a balancing factor λ_{bil} typically set between 0.5 and 1.

Finally, the smoothness regularisation term was introduced to penalise abrupt local variations in the predicted depth map:

$$\mathcal{L}_{smooth} = \frac{1}{N}\sum_{i}\left(\left|\partial_x \hat{d}_i\right| e^{-|\partial_x I_i|} + \left|\partial_y \hat{d}_i\right| e^{-|\partial_y I_i|}\right), \tag{8}$$

where $I_i$denotes the image intensity and ∂_x, $\partial_y$are the image gradients along the horizontal and vertical directions. This term encourages locally smooth depth predictions in homogeneous regions while preserving depth discontinuities near strong image edges.

According to the training logs, convergence was generally achieved well before the nominal limit of 100 epochs, often after approximately 30 epochs, with total training times ranging from about 4 to 10 hours on an NVIDIA RTX 4090. The loss curves shown in Figures 2 and 3 indicate stable convergence in both fine-tuning stages. In the first step, both training and validation

losses decrease rapidly and then progressively stabilise, with no clear evidence of severe overfitting. In the second step, convergence is even faster, as the training loss drops markedly during the first epochs and the validation AbsRel quickly reaches a low and stable level. Overall, these trends suggest that the first step provides the main correction of the domain gap, whereas the second step acts as a refinement stage, improving generalisation across mixed oblique and nadir views.

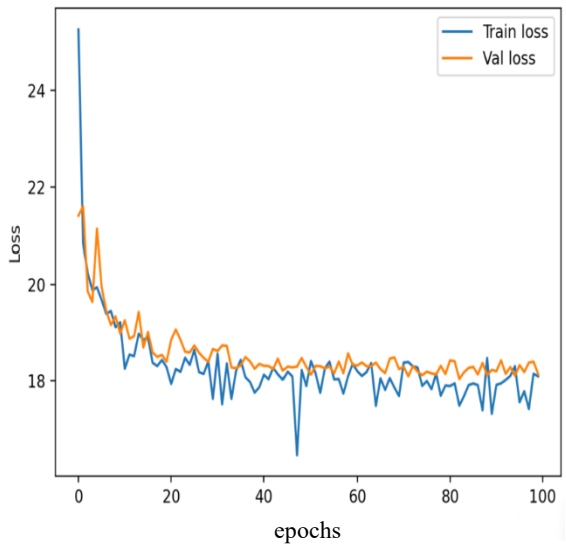

Figure 2. The trend of Train and Validation Loss of first step

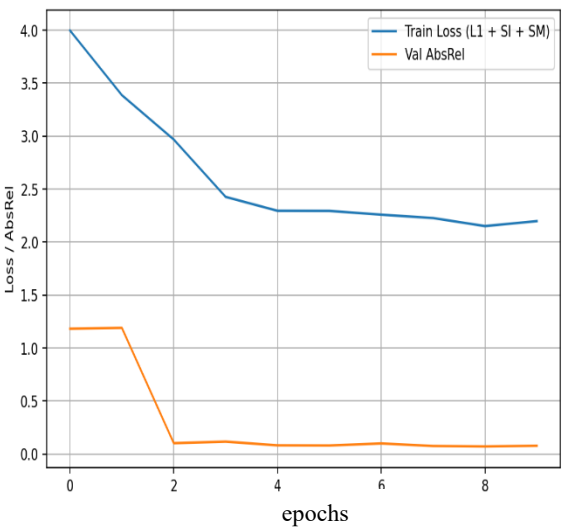

Figure 3. The trend of Train Loss and AbsRel of second step

Methodologically, this two-step design is relevant because it addresses the UAV adaptation problem in a realistic and operational way. Rather than redefining the entire network, it progressively specialises the depth-prediction components of Depth Anything V2, first under the most critical oblique conditions and then under a more heterogeneous viewpoint distribution. This makes the procedure more tractable for photogrammetric and heritage-documentation workflows, where the objective is not large-scale model development, but the controlled adaptation of an existing high-performing model to a specific acquisition domain.

4.3 Global synthetic results after fine-tuning

The fine-tuned model shows a substantial improvement (Table 4) with respect to the direct application of the metric Depth Anything V2 model before adaptation (Table 3). The most evident effect concerns the mean residuals, which move from large negative values, often on the order of several tens of metres, to values generally close to zero. This indicates that the fine-tuning procedure is effective not only in reducing dispersion, but also in correcting the strong systematic bias observed in the original metric baseline.

	set	m [m]	Std [m]	RMSE log	Abs Rel	Comments
A	All	-0.331	1.051	0.107	0.092	Bias strongly reduced; clear gain vs both baselines
A	N.	0.012	0.999	0.125	0.099	Bias corrected; still weaker than rescaled relative
A	O 45	-0.406	0.655	0.101	0.077	Strong gain; near-stable oblique response
A	O 60	-0.425	1.198	0.132	0.092	Major improvement over both baseline
D	All	0.093	0.579	0.015	0.008	Very large improvement; bias nearly removed
D	O 45	0.118	0.638	0.017	0.009	Strongest recovery in case D
D	O 60	0.053	0.535	0.014	0.009	Severe baseline failure largely corrected
C	All	0.065	0.592	0.028	0.009	Clear gain vs both baselines
C	N	0.012	0.999	0.025	0.015	Good correction; still some dispersion
C	O 60	0.096	0.710	0.038	0.033	Oblique subset markedly improved
G	All	0.521	1.401	0.026	0.015	Large gain; residual positive bias remains
G	N	0.495	1.952	0.045	0.031	Bias reduced, but nadir still critical
G	O 45	-0.056	1.087	0.015	0.014	Strong recovery; best subset in case G

Table 4 - Representative baseline results before fine-tuning, using Dav2 metric model

In comparison with the pre-fine-tuning metric model, the gains are consistent across all representative cases. For example, in case D-All, the mean residual improves from -25.7 m to 0.093 m, while Std decreases from 17.3 m to 0.579 m, RMSE log from 0.50 to 0.015, and Abs Rel from 0.69 to 0.008. Similar patterns are observed in cases C and G, confirming that the fine-tuning step largely removes the scale collapse and restores a much more balanced metric response.

The comparison with the rescaled relative model is more nuanced. In most subsets, the fine-tuned metric model yields a substantial improvement over the direct metric baseline and, in several cases, also over the rescaled relative baseline. However, the comparison is not uniform across all subsets and metrics. In particular, the nadir results do not support a simple generalisation, since the post-fine-tuning model shows a strong reduction of systematic bias, while the relative-rescaled baseline may still remain competitive for some error measures.

4.4 Some detailed analysis

A more detailed inspection of the post-fine-tuning residual maps indicates that the remaining errors are not randomly distributed and cannot be regarded as normally distributed. Rather, the residuals show clear spatial structure, with recurrent clustering over specific architectural elements and systematic border effects along sharp depth discontinuities. This confirms that the residual

component is still influenced by scene geometry and local radiometric variations, rather than being reduced to purely stochastic noise.

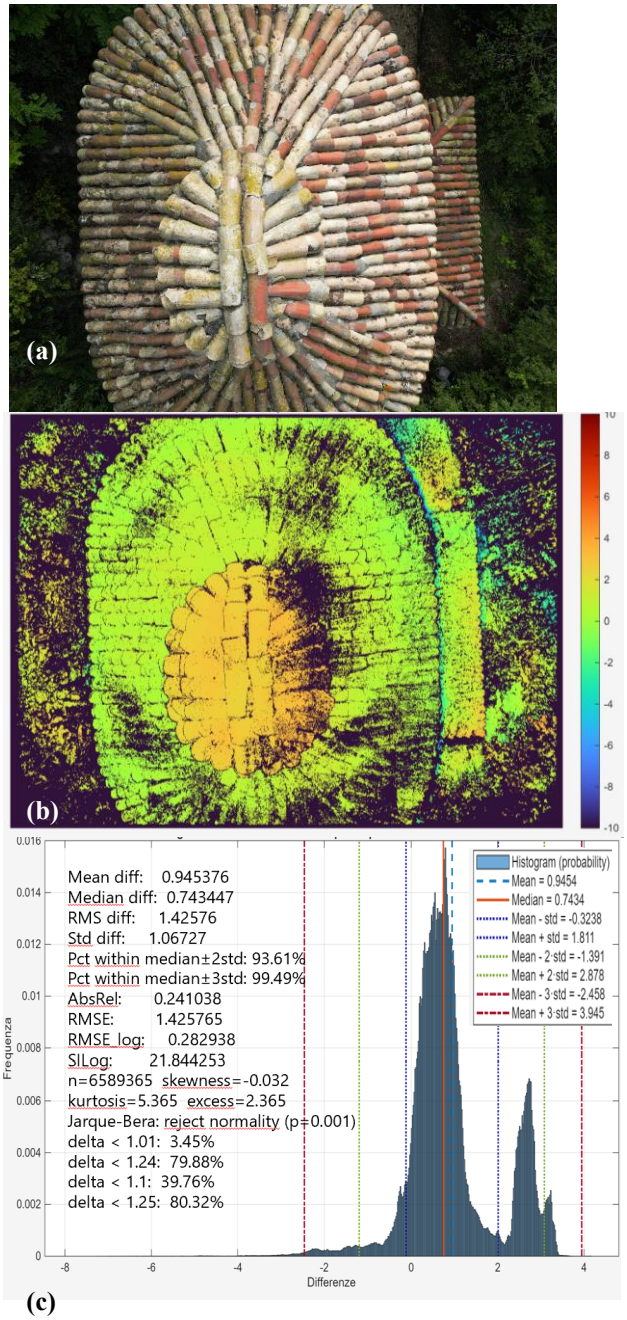

Figure 4. An example detailed analisys on case E: (a) original image, (b) dicrepancies[m], (c) histogram of discrepancies [m]

In particular, the detailed examples highlight two recurring patterns. First, pronounced edge-related residuals appear along roof outlines, façade boundaries, and other abrupt geometric transitions, indicating that object borders remain one of the most critical areas for prediction. Second, in the Sacro Monte di Crea case (Case E, Figure 4), the residual distribution shows a clear clustering associated with the different depth levels of the roof surfaces.

This suggests that the model still tends to organise part of the remaining error according to distinct geometric roof portions, rather than producing a uniform residual field. Such behaviour is consistent with the presence of multi-level and articulated roof morphologies, where local depth transitions are structured and not easily absorbed by the fine-tuned model.

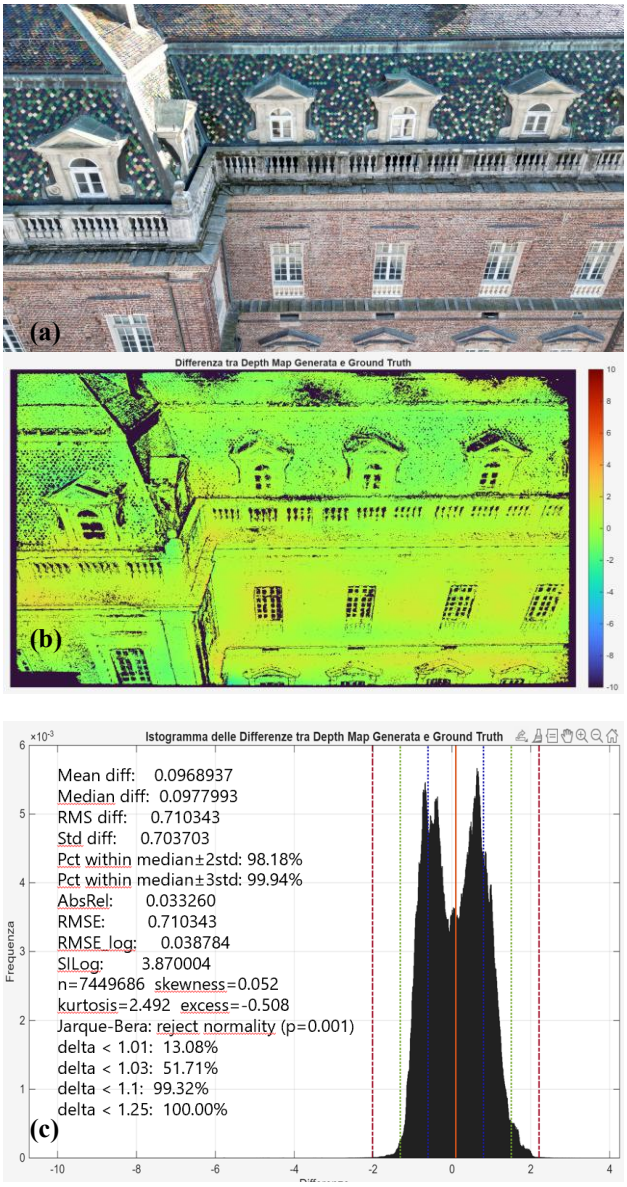

Figure 5. An example detailed analisys on case C: (a) original image, (b) dicrepancies[m], (c) histogram of discrepancies [m]

In the Venaria façade example (Figure 5), the residual field appears more controlled than in the roof-dominated cases, but the remaining errors are still concentrated around architectural discontinuities, such as roof-to-façade transitions, cornices, and thin vertical elements. This indicates that, even after adaptation, the model remains more stable on broader planar portions than on sharp edges and fine structural details. The façade case therefore confirms that the residual component is still partly driven by local geometry, especially where depth changes are abrupt or spatially compressed in the image.

In the urban-roof example (Timespan, Figure 6), the remaining defects are mainly associated with roof boundaries, ridges, and abrupt transitions between adjacent roof planes or between roofs and ground-level surfaces. Additional residual concentrations appear over portions characterised by heterogeneous roofing materials, colour differences, shadows, and nearby vegetation,

suggesting that the model is still sensitive to local radiometric variability and to small-scale geometric discontinuities. In this type of urban scene, the residual component therefore seems to be controlled less by broad planar surfaces than by edges, level changes, and heterogeneous roof textures

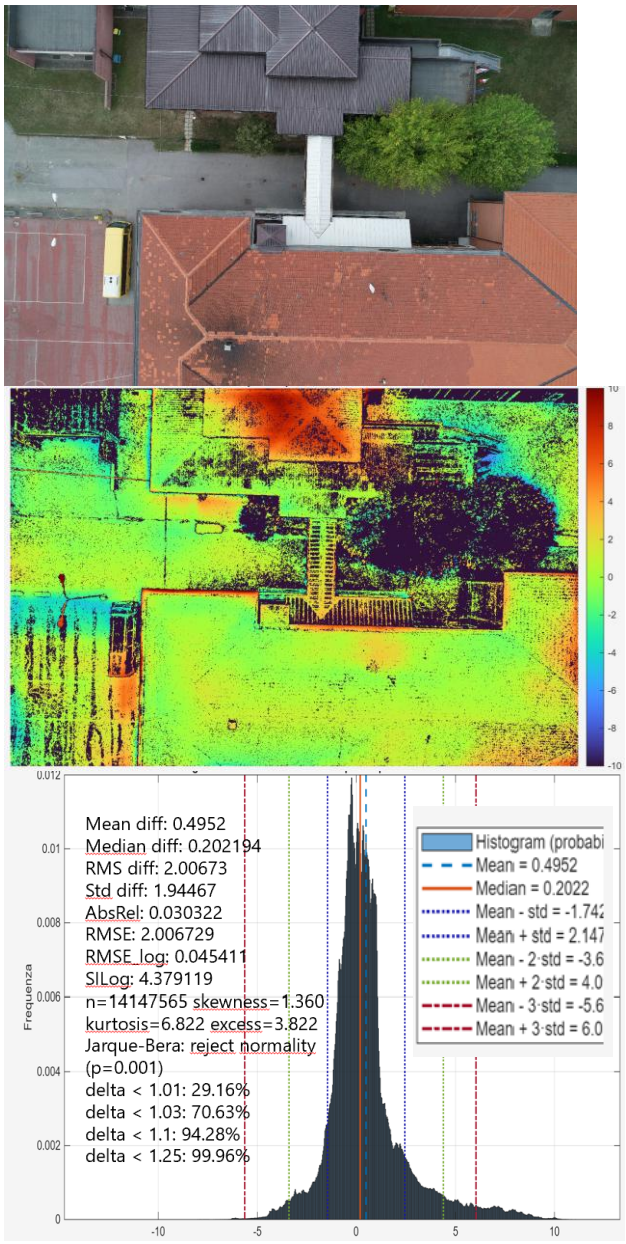

Figure 6. An example detailed analisys on case A: (a) original image, (b) dicrepancies[m], (c) histogram of discrepancies [m]

These detailed examples complement the global metrics by showing that the post-fine-tuning residuals are still non-normal, spatially clustered, and strongly affected by border effects. The adaptation substantially improves the global metric balance of the prediction, but residual artefacts persist near edges, thin elements, and multi-level roof structures.

Overall, the results show that the proposed two-step fine-tuning is effective in reducing bias, improving metric consistency, and substantially lowering dispersion and relative error. Nevertheless, the achieved accuracies are still insufficient for the most demanding metric applications, especially where precise geometric measurement is required. At the same time, the post-adaptation performance may already be useful for thematic applications, scene understanding, robotics, navigation, and collision-avoidance tasks, that is, in contexts where a reliable geometric trend is needed, but high metric precision is not essential.

5. Conclusions

This study shows that recent monocular depth estimation foundation models can provide useful geometric cues for UAV-based architectural heritage documentation, but only after a domain-specific adaptation process. The baseline experiments demonstrate that direct transfer from terrestrial vision to aerial heritage imagery is still affected by substantial metric inconsistency, scale distortion, and viewpoint-dependent errors, especially in oblique configurations and architecturally complex scenes. The proposed two-step fine-tuning strategy considerably improves metric balance, reduces systematic bias, and enhances the overall stability of the predictions across the benchmark, confirming that controlled retraining of the decoder and depth head is an effective adaptation strategy. At the same time, the detailed residual analysis highlights that the remaining errors are still spatially structured, non-normal, and concentrated near object boundaries, multi-level roofs, and radiometrically heterogeneous surfaces. The achieved results are therefore not yet sufficient for the most demanding metric applications, where high geometric accuracy is required, but they already indicate a promising potential for complementary uses in thematic interpretation, scene understanding, robotics, navigation, and collision-avoidance scenarios. In this sense, the proposed benchmark and experimental protocol provide an initial reference framework for assessing the role of monocular depth prediction as a complementary, rather than substitutive, source of information within heritage-oriented photogrammetric workflows.

Acknowledgments

The authors gratefully acknowledge SIFET (Italian Society of Photogrammetry and Surveying) for the images (cases D and L) acquired in the SUNRISE Summer School, co-funded by ISPRS. They also thank the Italian Fire and Rescue Service (Vigili del Fuoco, “Commando regionale del Piemonte”) for enabling the UAV acquisitions over the roofs of La Venaria Reale, as well as the students of the “Droni per il rilievo territoriale e architettonico” course for their contribution to the façade surveys at the same site. The student team DIRECT is also acknowledged for its active participation in several UAV field campaigns that supported the construction of this benchmark. The study was carried out within the framework of the collaboration with the FAIR – Future Artificial Intelligence Research project and received funding from the European Union – Next Generation EU (National Recovery and Resilience Plan – NRRP, Mission 4, Component 2, Investment 1.3, Decree No. 1555 of 11/10/2022, PE00000013). The content of this manuscript reflects only the views of the authors; neither the European Union nor the European Commission can be held responsible for them.

References

Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Mueller, M., 2023. ZoeDepth: Zero-shot transfer by combining relative and metric depth. arXiv:2302.12288. https://doi.org/10.48550/arXiv.2302.12288.

Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V., 2024. Depth Pro: Sharp monocular

metric depth in less than a second. arXiv:2410.02073. https://doi.org/10.48550/arXiv.2410.02073.

Cho, J., Min, D., Kim, Y., Sohn, K., 2021. DIML/CVL RGB-D dataset: 2M RGB-D images of natural indoor and outdoor scenes. arXiv preprint arXiv:2110.11590.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. In: Advances in Neural Information Processing Systems, 27, pp. 2366–2374.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The KITTI dataset. International Journal of Robotics Research, 32(11), 1231–1237.

Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K., 2024. Repurposing diffusion-based image generators for monocular depth estimation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 9492-9502.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks. In: Proceedings of the 4th International Conference on 3D Vision (3DV), pp. 239–248.

Nex, F., Stathopoulou, E. K., Remondino, F., Yang, M. Y., Madhuanand, L., Yogender, Y., Alsadik, B., Weinmann, M., Jutzi, B., Qin, R., 2024. UseGeo - A UAV-based multi-sensor dataset for geospatial research. ISPRS Open Journal of Photogrammetry and Remote Sensing, 13, 100070.

Nex, F., Stathopoulou, E. K., Remondino, F., Yang, M.Y., Madhuanand, L., Yogender, Y., Alsadik, B., Weinmann, M., Jutzi, B., Qin, R., 2024. UseGeo - A UAV-based multi-sensor dataset for geospatial research. ISPRS Open J. Photogramm. Remote Sens., 13, 100070. https://doi.org/10.1016/j.ophoto.2024.100070.

Piccinelli, L., Yang, Y.-H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F., 2024. UniDepth: Universal monocular metric depth estimation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 10106-10116.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 12179-12188.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12179–12188.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V., 2022. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Trans. Pattern Anal. Mach. Intell., 44(3), 1623-1637. https://doi.org/10.1109/TPAMI.2020.3019967.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V., 2022. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3), pp. 1623–1637.

Saxena, A., Chung, S. H., Ng, A. Y., 2005. Learning depth from single monocular images. In: Advances in Neural Information Processing Systems, 18, pp. 1161–1168.

Schöps, T., Schönberger, J. L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A., 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2538–2547.

Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from RGBD images. In: Computer Vision – ECCV 2012, Lecture Notes in Computer Science, 7576, 746–760. Springer, Berlin, Heidelberg.

Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J., 2025. MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 5261-5271.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H., 2024a. Depth Anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10371–10381.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024b. Depth Anything V2. Adv. Neural Inf. Process. Syst., 37. https://doi.org/10.52202/079017-0688.