Verso una stima di profondità metrica da immagini UAV per il patrimonio architettonico: il caso di Depth Anything V2

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

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

Abstract. Lo studio valuta Depth Anything V2 su oltre 3.000 immagini UAV di dieci siti del patrimonio architettonico, usando mappe fotogrammetriche come riferimento metrico. Il modello relativo, riscalato, e quello metrico mostrano errori legati al punto di vista; il secondo comprime sensibilmente le profondità. Un fine-tuning in due fasi, su viste oblique e nadirali, migliora la coerenza metrica, ma persistono errori ai bordi, sulle coperture a più livelli e nelle geometrie complesse.

Parole chiave: stima monoculare della profondità; UAV; patrimonio architettonico; fine-tuning; fotogrammetria; adattamento di dominio.

1 Introduzione

La stima monoculare della profondità (MDE) è passata da modelli convoluzionali specializzati per singoli scenari a modelli basati su transformer in grado di ricavare una struttura densa da una sola immagine RGB. MiDaS ha mostrato che l'addestramento su più dataset favorisce la generalizzazione senza ulteriore addestramento sul dominio di destinazione, mentre i dense prediction transformer hanno dimostrato l'utilità di integrare il contesto globale con i dettagli locali (Ranftl et al., 2022; Ranftl et al., 2021). Approcci più recenti hanno affrontato la stima metrica e il trasferimento tra domini: ZoeDepth, UniDepth, Marigold, Depth Anything V2, Depth Pro e MoGe (Bhat et al., 2023; Piccinelli et al., 2024; Ke et al., 2024; Yang et al., 2024; Bochkovskii et al., 2024; Wang et al., 2025).

Il patrimonio architettonico pone difficoltà specifiche a questi modelli. Le acquisizioni UAV oblique permettono di documentare le facciate e altre strutture verticali, integrando le viste nadirali nei rilievi di edifici storici (Aicardi et al., 2016; Masiero et al., 2019). In complessi architettonici estesi, la valutazione della qualità geometrica e l'integrazione di dati provenienti da sensori diversi costituiscono inoltre aspetti centrali del processo di documentazione (Chiabrando et al., 2019).Tessiture ripetitive, elementi decorativi sottili, discontinuità nette di profondità, ombre portate e

materiali diversi si accompagnano a frequenti passaggi tra coperture, facciate, vegetazione e cielo. Le acquisizioni da Uncrewed Aerial Vehicles (UAV) introducono inoltre viste nadirali e fortemente oblique, quote di presa e configurazioni della scena diverse da quelle delle immagini terrestri prevalenti nell'addestramento di molti sistemi MDE. Ottenere profondità metricamente attendibili è quindi più impegnativo rispetto alla generazione di di una ricostruzione visivamente plausibile. In questo studio la profondità stimata è valutata quale possibile fonte di informazione geometrica per la ricostruzione, la produzione di ortoimmagini, la premodellazione HBIM e la documentazione GIS, in integrazione con i prodotti del rilievo fotogrammetrico.

Depth Anything V2 è stato scelto perché offre modelli relativi e metrici nell'ambito di una stessa architettura, consente un adattamento controllato ed è disponibile pubblicamente (Yang et al., 2024). Il confronto tra le due configurazioni permette di distinguere la capacità di ordinare correttamente le profondità dal recupero della scala metrica nelle immagini UAV.

2 Stato dell'arte e requisiti dei dataset

I metodi pertinenti a questo studio affrontano aspetti differenti della stima della profondità da immagine singola. MiDaS apprende la profondità relativa mediante una supervisione eterogenea, ma non dispone di una scala metrica intrinseca (Ranftl et al., 2022). ZoeDepth combina stima relativa e metrica mediante componenti specifiche per dominio, mentre UniDepth mira a una previsione metrica più generalizzabile (Bhat et al., 2023; Piccinelli et al., 2024). Marigold impiega modelli di diffusione; Depth Anything V2 utilizza una supervisione teacher–student su larga scala; Depth Pro punta a contorni netti nella profondità metrica; MoGe stima una rappresentazione geometrica più articolata (Ke et al., 2024; Yang et al., 2024; Bochkovskii et al., 2024; Wang et al., 2025).

Nelle viste UAV del patrimonio costruito, la differenza rispetto ai dati di addestramento riguarda tanto la geometria quanto la radiometria. Nelle immagini terrestri, la profondità della scena è spesso organizzata attorno a strade, ambienti interni e linea dell'orizzonte. Nelle immagini aeree, le coperture possono occupare gran parte del fotogramma e l'asse ottico può essere rivolto verso il basso. Elementi a quote simili possono inoltre presentare brusche discontinuità locali. Un modello metrico addestrato su immagini terrestri può quindi comprimere l'intervallo delle profondità, anche quando conserva in parte il loro ordinamento locale.

Dataset metrici diffusi, tra cui KITTI, NYU Depth V2, ETH3D e DIML, non rappresentano pienamente i punti di vista e le distribuzioni di profondità propri dei rilievi UAV del patrimonio architettonico. Precedenti più vicini al presente studio sono ArchDepth, benchmark fotogrammetrico sintetico, e UseGeo, dataset multisensore acquisito da UAV; quest'ultimo è stato utilizzato anche per valutare la stima della profondità e la ricostruzione 3D da immagini (Welponer et al., 2022; Nex et al., 2024; Hermann et al., 2024). Il contributo qui presentato riunisce siti di interesse architettonico, viste nadirali e oblique, forti variazioni della quota relativa di volo e

riferimenti fotogrammetrici densi, predisposti per la calibrazione e l'adattamento supervisionato.

3 Il benchmark MDE4BH

L'attuale benchmark MDE4BH comprende oltre 3.000 immagini UAV relative a dieci casi di studio: isolati urbani, facciate monumentali, torri isolate, ville, una cupola, un esteso complesso architettonico e resti archeologici. Sono presenti acquisizioni nadirali, oblique e miste. Le quote relative di volo variano approssimativamente da 5 a 90 m; le immagini provengono da diverse piattaforme DJI, con differenti lunghezze focali e dimensioni del pixel. La Tabella 1 riporta la composizione dei casi di studio.

Le mappe di profondità di riferimento sono state prodotte in Agisoft Metashape mediante elaborazione a piena risoluzione e ricostruzione densa con impostazione Ultra High. Sono state mascherate le regioni di cielo, i vuoti, le superfici riflettenti e le altre aree non affidabili. Le mappe fotogrammetriche così ottenute forniscono il riferimento metrico per l'analisi dei residui, la riscalatura del modello relativo e il fine-tuning.

Tabella 1. Casi di studio del benchmark MDE4BH utilizzati per calibrazione, validazione e prove esterne.

ID	immagine	Drone/camera	n. immagini	Hrel [m]
(A) Timespan Museum (Scozia)		DJI Matrice M4E Focale 12.29 mm Pixel 3.36 mm	n. 80 nadirali incrociate n. 130 pitch ~ 60 ° n. 290 pitch ~ 45°	n. 150 ~ 20m n. 350 ~ 40 m
(B) Venaria Alto		DJI Matrice 350 Focale 35 mm Pixel 4.4 mm	n. 250 nadirali incrociate n. 180 pitch ~45°	~ 60m ~ 90 m
(C) Venaria Facciata		DJI Mini 3 Pro Focale 6.72 mm Pixel 2.4 mm	n.50 pitch ~ 90° n.25 pitch ~ 60°	18 m
(D) Torre Chianca		DJI Mavic 3M Focale 4.34 mm Pixel 3.3 mm	n. 40 pitch ~ 45° n. 55 pitch ~ 65°	n. 40 ~ 20 m n. 55 ~ 5 m
(E) Sacro Monte di Crea		DJI Mini 3 Pro Focale 6.72 mm Pixel 2.4 mm	n. 90 nadirali n. 95 pitch ~ 90° n. 30 pitch ~ 45° n. 115 pitch 60°-80°	5 m

ID	immagine	Drone/camera	n. immagini	Hrel [m]
(F) Villino Caprifoglio		DJI Spark Focale 4.4 mm Pixel 1.55 mm	n. 70 nadirali incrociate n.105 pitch 30°-70° n.25 pitch 90°	n. 130 ~ 25 m n.70 ~ 12 m
(G) Mappano		DJI Phantom 4 Pro	n. 195 nadirali incrociate n. 220 pitch 45°	80 m 60 m
(H) Cupola del Duomo di San Lorenzo		DJI Mini 3 Pro Focale 6.72 mm Pixel da 2.4 mm	n. 10 nadirali n.30 pitch 40°-60° n.90 60°-90°	13 m
(I) Castello del Valentino (To)		DJI Phantom 4 Pro Focale 8.8 mm Pixel size 2.41 µm	n. 209 Nadirali incrociate n. 67 inclinate 80 ° n. 205 inclinate 45°	n. 209 ~ 40 m n.272 ~ 60 m
(L) Bagno Arabo di Mezzagnone (RG)		DJIMini mavic 3 Focale 4.49 Pixel size 1.76	n. 229 Nadirali incorciate n. 139 inclinate 45°	n. 368 ~ 18 m

4 Protocollo sperimentale

Sono state valutate due configurazioni di riferimento. Nella prima, le previsioni del modello relativo Depth Anything V2 base sono state convertite in profondità metrica mediante regressione lineare sui soli pixel validi delle mappe fotogrammetriche. La scala e l'offset della trasformazione affine consentono di verificare quale parte della differenza tra i domini possa essere corretta con una calibrazione successiva. Nella seconda configurazione, il modello metrico outdoor è stato applicato direttamente alle immagini UAV. Il confronto impiega media e mediana dei residui, deviazione standard dei residui, radice dell'errore quadratico medio in scala logaritmica (RMSElog) ed

errore relativo assoluto (AbsRel); nelle valutazioni di dettaglio sono riportati anche SILog e le percentuali entro soglie δ.

4.1 Comportamento dei modelli di riferimento

Dopo la riscalatura, il modello relativo conserva un ordinamento utile delle profondità in alcune scene, soprattutto nelle viste urbane nadirali. Il modello metrico applicato direttamente mostra invece una marcata compressione della scala nelle geometrie di presa aeree: oggetti a circa 30–40 m dalla camera possono ricevere profondità previste di soli 5–10 m. La Tabella 2 riproduce alcune statistiche rappresentative dei modelli di riferimento.

A Mappano, le immagini nadirali sono relativamente stabili: la deviazione standard dei residui è nell'ordine di 0,5–1,3 m e gli errori relativi si collocano intorno all'1%; nelle viste oblique i valori peggiorano rispettivamente a circa 5,1–15,7 m e 5–10%. Per la facciata della Venaria, la dispersione dei residui varia da pochi decimetri a decine di metri secondo il sottoinsieme e l'angolo di vista. Strutture verticali isolate come la Torre Chianca risentono inoltre del punto di presa e della presenza del cielo. I risultati indicano che, oltre alla calibrazione della scala, occorre intervenire sulle regolarità geometriche apprese dal modello su scene terrestri.

Tabella 2. Risultati rappresentativi prima del fine-tuning; valori e segni dei residui riprendono il contributo originale (m = media, med = mediana, std = deviazione standard).

Caso	Sottoinsieme	m [m]	med [m]	std [m]	RMSElog	AbsRel
D rel. riscalato	Tutte	0,00	0,00	6,11	0,38	0,101
D rel. riscalato	Oblique 45°	0,00	0,00	4,61	0,12	0,068
D rel. riscalato	Oblique 60°	0,00	0,00	12,94	1,23	0,223
D metrico	Tutte	25,06	25,51	6,47	0,99	0,532
A rel. riscalato	Tutte	0,00	0,00	8,63	0,15	0,102
A metrico	Tutte	25,75	−24,90	17,27	0,50	0,692

4.2 Fine-tuning in due fasi

L'adattamento supervisionato è stato condotto in due fasi. La prima utilizza soprattutto immagini UAV oblique, che descrivono la geometria delle facciate e sono particolarmente critiche per il modello non adattato. La seconda riparte dal modello già aggiornato e introduce immagini oblique e nadirali per migliorare la generalizzazione rispetto al punto di vista. L'addestramento interessa principalmente il decoder e la componente di regressione della profondità; il *backbone* visivo resta congelato, o in larga parte congelato, per conservare le caratteristiche generali delle immagini e contenere il sovradattamento.

Per ciascuna fase sono state impiegate circa 1.000 immagini, con sottoinsiemi dedicati alla validazione e un insieme di prova costituito da immagini escluse

dall'addestramento. Gli ultimi due casi di studio sono stati esclusi interamente dall'addestramento. La funzione di costo combina errore L1 sulla profondità, termine logaritmico invariante alla scala e regolarizzazione della levigatezza, con pesi rispettivi di 1,0, 0,1 e 0,003. I tempi di addestramento riportati sono di circa 4–10 ore su NVIDIA RTX 4090. La funzione di costo tende a stabilizzarsi intorno a 30 epoche, prima del limite nominale di 100.

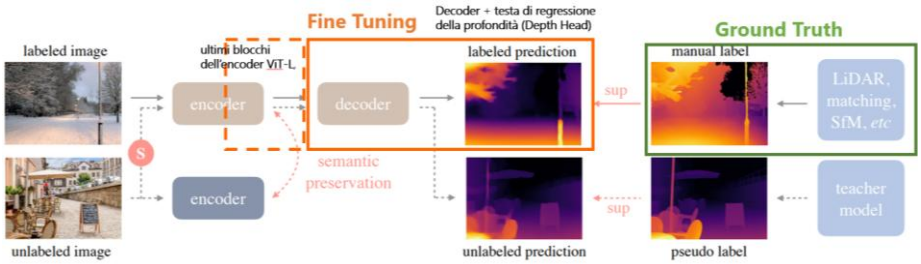

Fig. 1. Lo schema dell'algoritmo Depth Anything v2 con la parte oggetto di fine tuning

5 Risultati e discussione

Il fine-tuning migliora sensibilmente la coerenza metrica nel caso rappresentativo della Torre Chianca (D). Sull'insieme completo, la media dei residui è 0,093 m, la mediana 0,087 m e la deviazione standard 0,579 m; RMSElog è pari a 0,015 e AbsRel a 0,008. Valori simili si riscontrano nei sottoinsiemi obliqui a 45° e 60° (Tabella 3). Il miglioramento non è quindi limitato a un solo angolo di presa.

I confronti relativi al Sacro Monte di Crea, alla cupola di San Lorenzo e a Venaria Alto mostrano, dopo l'adattamento, residui medi prossimi a zero o nell'ordine di pochi decimetri. La deviazione standard diminuisce approssimativamente di un fattore compreso tra tre e quattro; gli intervalli riportati dopo l'addestramento sono rispettivamente circa 0,3–2,5 m, 1,2–2,7 m e 0,8–3,7 m, con AbsRel generalmente tra il 4% e l'11%.

Le analisi sui dettagli esclusi dall'addestramento mostrano sia i progressi sia i limiti dell'approccio. Per il dettaglio di Mappano si ottengono AbsRel 0,030, RMSE 2,01 m e il 99,96% dei pixel validi entro $\delta < 1,25$. Il dettaglio della Venaria raggiunge AbsRel 0,033 e RMSE 0,71 m. Il Sacro Monte di Crea resta più difficile: AbsRel è pari a 0,241, i residui si distribuiscono in gruppi distinti e persistono errori sulle coperture a quote differenti. Anche la non normalità dei residui, gli effetti ai bordi, le variazioni cromatiche e la geometria complessa rimangono problematici. La profondità stimata può pertanto sostenere interpretazioni geometriche rapide, integrazioni circoscritte delle lacune e procedure ibride, nel rispetto dei requisiti di accuratezza della specifica applicazione.

Infatti, le valutazioni di dettaglio, riportate a titolo d'esempio nelle fig 2 e fig. 3, mostrano comportamenti diversi nei due siti. A Mappano (fig. 2), il residuo medio è pari a 0,495 m, la deviazione standard a 1,945 m e l'AbsRel a 0,030; il 99,96% dei pixel

validi rientra nella soglia $\delta < 1{,}25$. La distribuzione dei residui resta tuttavia asimmetrica e la mappa evidenzia errori localizzati lungo i bordi e su alcune coperture.

Tabella 3. Risultati selezionati dopo l'adattamento per il caso D e per dettagli esclusi dall'addestramento (met. = metrico, fn = fine-tuning).

Caso	Sottoinsieme	Media [m]	Mediana [m]	Dev. std. [m]	RMSElog	AbsRel
D metrico fn	Tutte	0,093	0,087	0,579	0,015	0,008
D metrico fn	Oblique 45°	0,118	0,144	0,638	0,017	0,009
D metrico fn	Oblique 60°	0,053	0,054	0,535	0,014	0,009
G dettaglio	Escluso dall'addestr.	0,495	0,202	1,945	0,045	0,030
C dettaglio	Escluso dall'addestr.	0,097	0,098	0,704	0,039	0,033
E dettaglio	Escluso dall'addestr.	0,945	0,743	1,067	0,283	0,241

Al Sacro Monte di Crea (un esempio in fig. 3), pur essendo inferiore la deviazione standard (1,067 m), l'AbsRel sale a 0,241 e la quota di pixel entro $\delta < 1{,}25$ scende all'80,32%. L'istogramma presenta un secondo gruppo distinto di residui, associato nella mappa ad alcune porzioni della copertura a quote differenti. La sola dispersione complessiva non descrive quindi adeguatamente la qualità della stima: occorre considerare anche l'errore relativo e la distribuzione spaziale dei residui, soprattutto nelle geometrie architettoniche più articolate.

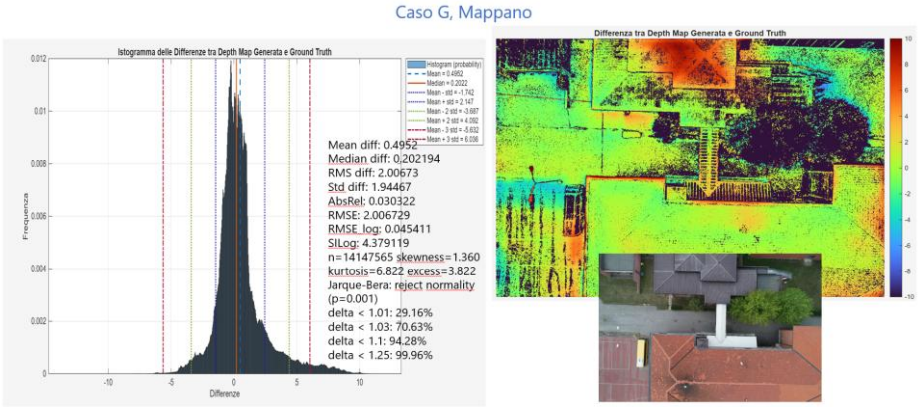

Fig. 2. Una caso esemplificativo delle valutazione di dettaglio nel caso di Mappano

6 Conclusioni

I modelli Depth Anything V2 addestrati prevalentemente su immagini terrestri non possono essere trasferiti direttamente alle immagini UAV del patrimonio architettonico quando è richiesta affidabilità metrica. Gli errori dei modelli di riferimento dipendono

fortemente dal punto di vista e dalla scala della scena; il modello metrico outdoor è particolarmente soggetto alla compressione delle profondità. Le mappe fotogrammetriche mascherate dei casi MDE4BH forniscono i dati necessari per valutare e correggere tale differenza.

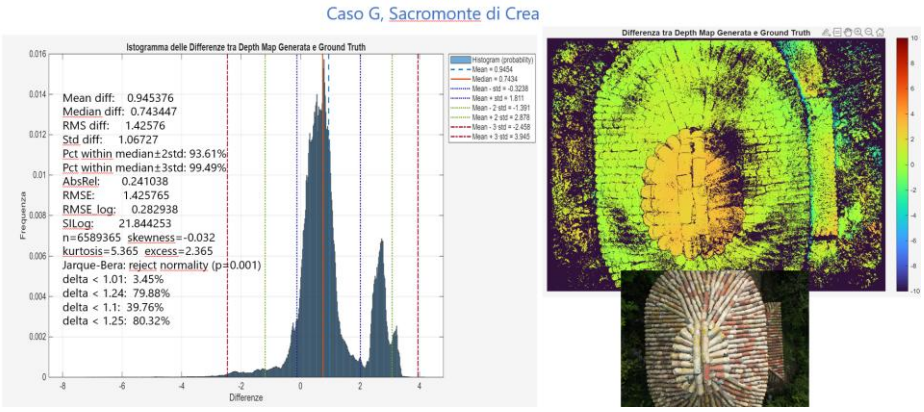

Fig. 3. Una caso esemplificativo delle valutazione di dettaglio nel caso del tetto di una cappella Sacro Monte di Crea

Un fine-tuning mirato in due fasi recupera una parte rilevante della coerenza metrica, anche in diverse viste oblique. La struttura dei residui sulle coperture complesse, ai bordi delle immagini e nei casi più difficili esclusi dall'addestramento limita tuttavia l'impiego attuale del metodo nei prodotti di rilievo ad alta precisione. La stima monoculare adattata può apportare informazione geometrica ai flussi di documentazione del patrimonio, in integrazione con la ricostruzione fotogrammetrica rigorosa.

Ringraziamenti

Lo studio è stato condotto nell'ambito del progetto FAIR — Future Artificial Intelligence Research e ha ricevuto finanziamenti dall'Unione europea — Next Generation EU (Piano Nazionale di Ripresa e Resilienza — PNRR, Missione 4, Componente 2, Investimento 1.3 — D.D. 1555 dell'11/10/2022, PE00000013). Il contenuto del contributo riflette esclusivamente le opinioni degli autori. Si ringra altresì il team studentesco DIRECT che ha collaborato con le acquisizioni di alcuni casi trattati, i Vigili del Fuoco sezione APR di Torino per la acquisizioni sulla Reggia di Venaria, Francesca Matrone, Marco Zerbinatti, Edward Borgogno, Beatrice Tanduo per le acquisizioni al Sacro Monte di Crea.

Riferimenti bibliografici

1. Aicardi, I., Chiabrando, F., Grasso, N., Lingua, A. M., Noardo, F., Spanò, A. (2016). UAV Photogrammetry with Oblique Images: First Analysis on Data Acquisition and Processing. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLI-B1, 835–842. DOI: 10.5194/isprs-archives-XLI-B1-835-2016.
2. 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.
3. 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.
4. Hermann, M., Weinmann, M., Nex, F., Stathopoulou, E.K., Remondino, F., Jutzi, B., Ruf, B. (2024). Depth estimation and 3D reconstruction from UAV-borne imagery: Evaluation on the UseGeo dataset. ISPRS Open Journal of Photogrammetry and Remote Sensing, 13, 100065. https://doi.org/10.1016/j.ophoto.2024.100065.
5. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K. (2024). Repurposing diffusion-based image generators for monocular depth estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9492–9502.
6. Masiero, A., Chiabrando, F., Lingua, A. M., Marino, B. G., Fissore, F., Guarnieri, A., Vettore, A. (2019). 3D Modeling of Girifalco Fortress. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W9, 473–478. DOI: 10.5194/isprs-archives-XLII-2-W9-473-2019.
7. 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. https://doi.org/10.1016/j.ophoto.2024.100070.
8. Piccinelli, L., Yang, Y.-H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F. (2024). UniDepth: Universal monocular metric depth estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10106–10116.
9. Ranftl, R., Bochkovskiy, A., Koltun, V. (2021). Vision transformers for dense prediction. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 12179–12188.
10. 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), 1623–1637. https://doi.org/10.1109/TPAMI.2020.3019967.
11. 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. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5261–5271.
12. Welponer, M., Stathopoulou, E.K., Remondino, F. (2022). Monocular depth prediction in photogrammetric applications. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B2-2022, 469–476. https://doi.org/10.5194/isprs-archives-XLIII-B2-2022-469-2022.
13. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H. (2024). Depth Anything V2. Advances in Neural Information Processing Systems, 37. https://doi.org/10.52202/079017-0688.