Benchmarking Metric Depth for Construction Perception: Industrial Rebar Scenes with Exact Synthetic Ground Truth

Salman Hajizada
Carnegie Mellon University in Qatar
Email: shajizad@andrew.cmu.edu

Diram Tabaa
Carnegie Mellon University in Qatar
Email: dtabaa@andrew.cmu.edu

Gianni Di Caro
Carnegie Mellon University in Qatar
Email: gdicaro@andrew.cmu.edu

***Abstract*— Construction inspection needs depth in meters at centimeter tolerance, but driving and indoor benchmarks test neither thin reinforcing bars across working distances nor monocular models against calibrated geometry on the same pixels. We evaluate four monocular metric models alongside stereo and ArUco-scaled COLMAP on rebar scenes. On a physical sequence, swapping the pseudo-reference moves reported monocular AbsRel by up to $1.5\times$ for the best model and valid support from 80% to 44%, so we build an exact Blender benchmark: a 100 mm lattice, a stirrup pile, and a 231-bar pier at 0.22–17 m. Stereo and anchored multiview hold 1.1–2.0 cm RMSE at close range; native monocular RMSE reaches 20–151 cm. An oracle per-view scale cuts close-range error to 0.029–0.061 AbsRel, so most native error is a global scale bias, not lost structure. That correction reverses with distance, from $s \approx 0.16$–0.63 on the lattice to 3.16–12.58 on the pier, where 0.147–0.186 AbsRel of shape error survives scaling. Monocular metric depth is thus better used as a source of close-range relative structure than as a metric sensor across construction distances. The evaluation set and evaluation code are available at `https://huggingface.co/datasets/sal0-h/industrial-rebar-metric-depth`.**

I. INTRODUCTION

Construction robotics and industrial inspection increasingly depend on image-based 3D perception [1]–[4], and the quantities at stake are physical: reinforcement spacing, concrete cover, and defect depth must be read in meters to within about a centimeter before being checked against specification. Which sensor supplies those meters is the choice we reconsider. LiDAR and laser return range directly at millimeter-to-centimeter accuracy, but are expensive, heavy, slow to deploy, and sparse on the geometry inspection cares about, since a narrow bar against open space slips between beams. Calibrated stereo is cheaper and dense and fixes scale from its baseline, yet triangulation error grows with the square of distance: our stereo holds to a centimeter or two on the close lattice and degrades to meters on the distant pier (Sec. III).

A single ordinary camera avoids both the scanner's bulk and the rig's range limit. Already mounted on robots, drones, and site cameras, it could serve appearance and measurement at once. The obstacle is that a frame carries no reference length, since a small nearby object and a larger distant one project alike. Monocular models must recover absolute scale from learned image-to-depth statistics and supplied or inferred intrinsics, and that scale need not hold as working distance changes. Industrial capture sharpens this, with thin bars, self-occlusion in dense assemblies, and working distances from centimeters to meters, all outside these models' training.

Zero-shot metric models are built for exactly this, and report strong results on driving and indoor benchmarks, chiefly KITTI and NYU-style sets [5], [6]. Those scores do not settle the industrial question. The headline criterion δ_1 counts pixels with depths $\max(\hat{z}/z, z/\hat{z}) < 1.25$, a tolerance far looser than inspection allows, and many protocols align predictions to ground truth before scoring, which removes native scale from the result. We use KITTI only as an implementation check against published behavior (Sec. III).

Industrial evaluations stop short as well. Padkan et al. [7] score current metric models on isolated tabletop objects by aligned cloud-to-mesh error; assemblies go untested, stereo and multi-view stereo are never scored on the same pixels, and scale bias is never separated from relative structure. We therefore ask whether monocular metric models reach the precision industrial measurement requires, and where they fall short, whether the deficit lies in scale or in shape.

Answering that requires a depth reference trustworthy on the reinforcement itself. Dense surveyed depth over rebar is impractical, so on a physical lattice we fall back on two *pseudo*-references: FoundationStereo [8] with measured baseline and focal length, and COLMAP multi-view stereo [9], [10] scaled by ArUco fiducials [11] of known size. Both are estimates, and swapping between them changes reported monocular error by up to $1.5\times$ while substantially changing valid support, so such a score cannot separate model error from reference error. We therefore build a Blender [12] benchmark spanning multiple reinforcement geometries and working distances, with exact optical-axis depth and structure masks, scoring only masked reinforcement and applying an oracle per-view scale $s = \mathrm{med}(\mathrm{GT})/\mathrm{med}(\mathrm{pred})$ that separates native scale error from residual shape error.

Related work. Relative-depth models trained on mixed data under scale- and shift-invariant objectives transfer broadly but discard absolute scale [13]–[15]; metric behavior is a separate specialization. The metric models we test differ mainly in where scale comes from. Metric3D v2 [16] normalizes training into a canonical camera space and converts back with the camera model, and Depth Anything 3 [17] follows the same canonical-camera route. UniDepth v2 [18] predicts an internal camera representation jointly with pseudo-spherical geometry and needs no supplied intrinsics, while

Depth Pro [19] converts a canonical inverse-depth map to meters with a focal length that may be supplied or estimated. All report zero-shot accuracy on indoor, outdoor, and driving benchmarks; none reports how native scale behaves at sub-meter distance on metal reinforcement, and aggregate scores do not separate global scale error from relative structure.

Under industrial conditions, Padkan et al. [7] score Depth Pro, Depth Anything V2, and Metric3D v2 on five NeRFBK tabletop objects (6–30 cm) by cloud-to-mesh RMSE after ICP, finding Depth Pro strongest but industrial objects still difficult; their scenario is isolated parts, their real-object reference is itself a reconstruction, and stereo and SfM appear as fallbacks rather than baselines scored on the same pixels. Midwinter et al. [4] fine-tune monocular models on an in-situ LiDAR RGB-D corpus of railway overpass bridges for spalling extent, with accumulated LiDAR rather than exact depth as reference. Neither scores structural foreground against exact per-pixel depth, and neither reports calibrated stereo or fiducial-scaled multi-view stereo as depth-map baselines on those images.

Contributions. This paper makes three contributions:

1) An exact synthetic benchmark for reinforcement geometry: a Blender generator supplying optical-axis metric depth, structure masks, known camera geometry, and controlled viewpoints across scene types and distances.
2) Evidence that physical scores are reference-dependent: swapping FoundationStereo for ArUco-scaled COLMAP changes reported monocular error and valid support, so a pseudo-reference score cannot separate model error from reference error.
3) An exact-ground-truth diagnosis separating monocular metric scale from shape: stereo and multiview stay centimeter-accurate while native monocular error is one to two orders larger, an oracle per-view scale recovers close range to 0.029–0.061 AbsRel on the lattice, and the required correction reverses direction at long range where 0.147–0.186 AbsRel of shape error remains.

II. EVALUATION SETUP

A. Model selection

We evaluate four off-the-shelf monocular metric models using official checkpoints, with no fine-tuning, chosen for distinct roles in the current zero-shot landscape. Metric3D v2 [16] (`metric3d_vit_small`) is the canonical-camera baseline and receives the known intrinsics. UniDepth v2 [18] (`unidepth-v2-vitl14`) is the contrasting RGB-only case: it infers camera geometry and is not given intrinsics. Depth Pro [19] is included for its high-resolution, sharp-boundary design, which matters for thin bars; we supply the known focal length used to convert its canonical inverse-depth prediction to meters. Depth Anything 3 Metric [17] (`DA3METRIC-LARGE`) is given the known intrinsics; we convert its output to meters with the supplied focal length. Metric3D v2 and Depth Pro overlap directly with [7]; that study used Depth Anything V2, an earlier generation. Geometric baselines are FoundationStereo [8] and ArUco-scaled COLMAP multi-view stereo.

Physical lattice Disordered stirrups 231-bar pier

Fig. 1: Evaluation scenes: physical stereo lattice (left), disordered synthetic stirrups (center), and 231-bar pier (right).

B. Metrics and the scale diagnostic

We report absolute relative error (AbsRel), root-mean-square error (RMSE) in centimeters, and δ_1, the fraction of pixels with $\max(\hat{z}/z, z/\hat{z}) < 1.25$. Mean absolute error (MAE) measures average absolute deviation in physical units. We use RMSE because its squared residuals emphasize large depth errors that can distort bar spacing and position measurements. AbsRel permits comparison from the sub-meter lattice to the 8–17 m pier. On synthetic scenes, metrics use only masked target pixels with valid ground truth and prediction, macro-averaged over views, keeping background out of the score; masks cover 4.5–7.2% of the frame. Native evaluation uses uncorrected output. We also apply an oracle per-view scale $s = \mathrm{med}(\mathrm{GT})/\mathrm{med}(\mathrm{pred})$, independently on each image, in the scale-invariant tradition of [6], [13]. Each view gets its own scalar. Tables report the median of these per-view factors. AbsRel after the same scaling is AbsRel_s. The gap between native and scaled error separates a global multiplicative scale mismatch from wrong relative structure, and s measures the required per-view correction.

C. Data

Physical capture. A first-generation Stereolabs ZED (B=120 mm) recorded 1,001 stereo pairs of a 3D-printed 100×100×100 mm rebar-like lattice (alternating 7 mm and 4 mm rods on a 20 mm pitch) on an A3 ArUco board (Fig. 1, left). FoundationStereo is given the rectified stereo pair. COLMAP multi-view stereo [9], [10] on the left images produces per-view dense geometric depth. ArUco Umeyama alignment to triangulated marker corners supplies the global metric scale that multiplies that depth. Both are estimates, not survey ground truth. Physical monocular scores use the full frame: pixels where that pseudo-reference is valid and the prediction is finite and positive in the evaluation range. Coverage is 80.3% against FoundationStereo and 44.4% against COLMAP, identical across monocular models.

Synthetic generation. Lattice, pile, and pier are three geometry sources through one Blender 4.5 stereo renderer [12]: an imported CAD lattice, physics-settled stirrup instances, and a fixed pier assembly. The generator controls assembly, lighting, camera paths, optics (ideal pinhole versus factory ZED), stereo extrinsics, and per-view export. Poses are deterministic rings or truncated-sphere domes aimed at the assembly. Ideal optics use a shared pinhole K with parallel cameras; ZED optics use the factory K and distortion, rendered pinhole then warped onto the 1280×720 raw grid. Left/right extrinsics are exact and known (B=120 mm). Raw outputs are co-registered on that grid: left/right RGB, optical-axis depth in

meters, a binary structure mask, poses, and K. For ZED-optics stereo evaluation, RGB, depth, and masks are rectified together onto the evaluation grid. Masks are a rebar emission pass; a one-pixel rim is excluded at evaluation, so scores use interior reinforcement only.

Scenes. Fig. 1 shows the three evaluation geometries. The synthetic lattice twins the printed 100 mm mockup. Camera protocols run from a 100-view orbit to 504-view ideal and ZED-optics domes at 0.22–0.43 m. A physics-settled pile of six 10 mm overlapping-hook stirrups (280×200 mm) is captured at 0.50–0.90 m. A 231-bar pier (extent $4.80\times4.80\times6.67$ m) is captured at 8–17 m. Stirrup families and stacking follow [3]. Lab scenes include the A3 board in the RGB; the pier has no fiducial, so COLMAP there is median-anchored to ground truth rather than metric.

Scale anchoring. On the 504-view ZED-rectified lattice we replace the ArUco scale without re-running the reconstruction. A known-depth view is an exact Blender depth map: $r = \mathrm{med}(z_{\mathrm{GT}}/z_{\mathrm{COLMAP}})$ on the same interior-mask support as the synthetic scores. The reconstruction scale is the median of those ratios for anchor counts spanning k=1 to 300 views. At each k, 24 independent draws sample views without replacement inside a trial. A second run uses 12 markers instead of 6 on the A3 board and re-runs COLMAP.

III. RESULTS

A. KITTI check

On the 697-image Eigen split [5], [6], Metric3D v2 gave the lowest AbsRel (0.0926) and UniDepth v2 the lowest RMSE (3.96 m) with the highest δ_1 (0.926); Depth Anything 3 Metric and Depth Pro followed at 0.1196 and 0.1386 AbsRel. These match published behavior.

B. Synthetic scenes

Table I summarizes the three exact-GT scenes (504 views each; masked reinforcement). On the close-range lattice and disordered pile, FoundationStereo and ArUco-scaled COLMAP reach 1.13–2.03 cm RMSE with $\delta_1 \approx 0.99$. The best native monocular results are substantially worse: 20.09 cm for Metric3D v2 on the lattice and 15.46 cm for Depth Pro on the pile.

Fig. 2 plots native versus oracle-scaled AbsRel on all three scenes. On the lattice, native monocular AbsRel spans 0.750–5.883, but falls to 0.029–0.061 after scaling. The ranking also changes: Metric3D v2 is best natively but worst after scaling, while UniDepth v2 moves from worst natively to nearly best. Native metric accuracy and recovered relative structure are therefore distinct properties.

The required correction changes strongly with working distance. On the lattice (0.22–0.43 m), all four monocular models have $s < 1$ (0.16–0.63), placing the structure too far away. On the pile (0.50–0.90 m), s moves toward and across one (0.72–1.36). On the pier (8–17 m), all four have $s > 1$ (3.16–12.58), placing the structure too near. Native AbsRel also improves from lattice to pile for all four models even though oracle-scaled error does not generally improve,

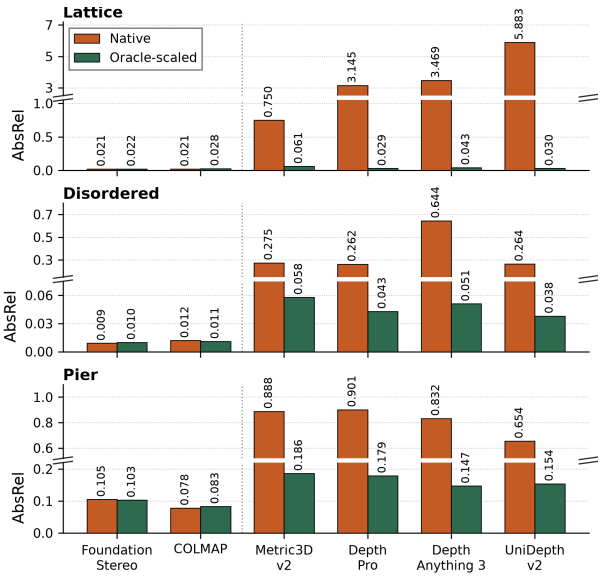

Fig. 2: Native vs. oracle-scaled AbsRel on masked reinforcement for the lattice, pile, and pier.

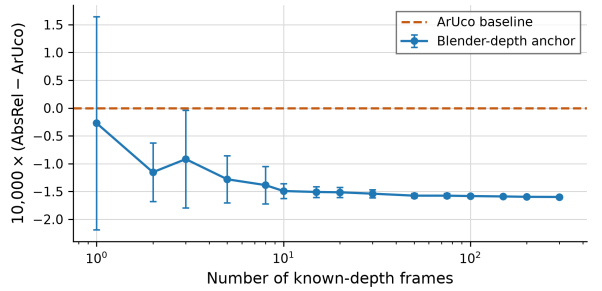

Fig. 3: Change in COLMAP AbsRel from ArUco versus number of known-depth views on the 504-view lattice (24 trials).

indicating that much of this gain comes from reduced scale bias rather than improved relative structure.

For FoundationStereo and COLMAP, $s \approx 1$ on the close-range scenes and oracle scaling barely changes their errors, confirming that their residual error is geometric rather than scale-driven. Monocular behavior differs (Fig. 2): although per-view scaling reduces close-range AbsRel to 0.029–0.061 on the lattice and 0.038–0.058 on the pile, it leaves 0.147–0.186 on the pier. Thus, close-range monocular error is largely multiplicative scale error, whereas at long range substantial residual structure error remains. The geometric baselines retain lower pier AbsRel of 0.078–0.105, 2.07–3.13 m RMSE.

C. How much anchoring multiview needs

On the 504-view lattice, one known-depth view recovers the mean COLMAP global scale to within 0.15% of ArUco (24 trials). Adding views (k=1 to 300) changes mean AbsRel only in the fourth decimal and mainly reduces scale variability (Fig. 3). Re-running COLMAP with 12 markers instead of 6 moves RMSE from 1.60 cm to 1.69 cm. Additional known-depth views mainly improve scale repeatability, while the 12-marker reconstruction does not improve depth error.

TABLE I: Synthetic scenes with exact Blender ground truth (504 views, masked reinforcement). †Median-anchored (no fiducial).

	Lattice (0.22–0.43 m)				**Disordered** (0.50–0.90 m)				**Pier** (8–17 m)			
Method	RMSE↓	AbsRel↓	AbsRel$_s$↓	s	RMSE↓	AbsRel↓	AbsRel$_s$↓	s	RMSE↓	AbsRel↓	AbsRel$_s$↓	s
FoundationStereo	**1.13**	0.021	0.022	1.00	**1.53**	0.009	0.010	1.00	312.8	0.105	0.103	0.97
COLMAP+ArUco	1.60	0.021	0.028	1.01	2.03	0.012	0.011	1.00	**206.7**†	0.078	0.083	0.98
Metric3D v2	20.09	0.750	0.061	0.63	20.46	0.275	0.058	1.36	1143	0.888	0.186	12.58
Depth Pro	91.19	3.145	0.029	0.25	15.46	0.262	0.043	0.92	1151	0.901	0.179	12.17
Depth Anything 3 Metric	96.93	3.469	0.043	0.23	35.31	0.644	0.051	0.72	1073	0.832	0.147	6.60
UniDepth v2	151.5	5.883	0.030	0.16	18.52	0.264	0.038	1.25	869.4	0.654	0.154	3.16

TABLE II: Physical ZED sequence against two pseudo-references.

	vs. FoundationStereo [8]				**vs. COLMAP** [9], [10]			
Method	RMSE↓	AbsRel↓	AbsRel$_s$↓	s	RMSE↓	AbsRel↓	AbsRel$_s$↓	s
Depth Anything 3 Metric [17]	**31.6**	**0.62**	0.201	0.64	**57.2**	**0.95**	0.265	0.54
Depth Pro [19]	42.8	0.90	0.185	0.61	59.7	0.98	0.231	0.60
UniDepth v2 [18]	45.8	0.98	**0.152**	0.56	63.4	1.07	**0.186**	0.56
Metric3D v2 [16]	58.7	1.40	0.253	0.49	72.0	1.54	0.328	0.49

D. Physical evaluation is reference-dependent

Where Table I scores FoundationStereo and COLMAP against exact depth, Table II uses them as competing full-frame pseudo-references for the monocular models. The ranking stays the same under both references, but the magnitudes do not. Depth Anything 3 Metric is best under both, at AbsRel 0.62 against FoundationStereo and 0.95 against COLMAP, a $1.5\times$ change for that model. The other models move less. Valid coverage also differs (80.3% against FoundationStereo vs. 44.4% against COLMAP). A physical score therefore mixes model error with the choice of reference.

IV. DISCUSSION AND LIMITATIONS

At close range the monocular models recover useful relative structure but place it at the wrong metric scale (Fig. 2). The required correction depends on the scene regime and reverses between the near lattice and the distant pier, so no single global rescaling makes these predictors metric across the tested distances. That reversal is consistent with extrapolation away from the intermediate ranges these models see in training: sub-meter structure is pushed too far and 8–17 m structure pulled too near. At long range the error surviving oracle scaling also grows, and scale correction alone no longer suffices. The geometric baselines behave differently. Calibrated stereo and ArUco-anchored multiview keep the correct scale on both close-range scenes, including the disordered pile, and their residual error is geometric rather than scale-driven. They are not immune to distance: both exceed 2 m RMSE on the pier, though their AbsRel stays well below the monocular models'. COLMAP's scale is a single global scalar, which one known-depth view recovers to within 0.15% of ArUco; neither additional anchor views nor doubling the markers improves depth error (Fig. 3). For construction metrology these results favor calibrated or externally anchored geometry where absolute measurement is required, and treat monocular depth as a source of relative structure rather than a sensor across working distances. Researchers can use the released scenes and masks to compare new models under fixed conditions, and use native versus oracle-scaled scores to distinguish scale mismatch from residual structure error.

Limitations. The centimeter conclusions rest on synthetic ground truth, since the physical sequence lacks surveyed per-bar depth. Appearance covers one controlled lab scene and synthetic counterparts, without varying illumination, metallic reflection, shadows, clutter, occlusion or weathered surfaces. The pier has no fiducial, so COLMAP there is median-anchored to ground truth and is not an independent metric result at that range. We fine-tune no model, and evaluation uses AbsRel, RMSE, and δ_1 rather than a boundary, surface-completeness, or thin-structure metric. Future work includes active view planning and a synthetic LiDAR baseline on the existing meshes.

REFERENCES

[1] A. D. Nair, J. Kindle, P. Levchev, and D. Scaramuzza, "Hilti SLAM Challenge 2023: Benchmarking single + multi-session SLAM across sensor constellations in construction," in *ICRA Workshop on Future of Construction*, 2024.

[2] H. Song *et al.*, "The city that never settles: Simulation-based LiDAR dataset for long-term place recognition under extreme structural changes," in *ICRA Workshop on Future of Construction*, 2025.

[3] T. Sun, B. Han, S. Rusinkiewicz, and Y. Shao, "Rebar grasp detection using a synthetic model generator and domain randomization," *Automation in Construction*, vol. 176, p. 106252, 2025.

[4] M. Midwinter, Z. A. Al-Sabbag, R. Bajaj, and C. M. Yeum, "Learning monocular depth estimation for defect measurement from civil RGB-D dataset," *Structural Health Monitoring*, vol. 25, pp. 2055–2070, 2026.

[5] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *CVPR*, 2012.

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

[7] N. Padkan, Z. Yan, and F. Remondino, "Evaluating monocular depth estimation methods on industrial objects," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XLVIII-1/W6, pp. 175–181, 2025.

[8] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "FoundationStereo: Zero-shot stereo matching," in *CVPR*, 2025.

[9] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *CVPR*, 2016.

[10] J. Schönberger *et al.*, "Pixelwise view selection for unstructured multi-view stereo," in *ECCV*, 2016.

[11] S. Garrido-Jurado *et al.*, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

[12] Blender Foundation, "Blender 4.5 LTS," Software, 2025.

[13] R. Ranftl *et al.*, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, 2022.

[14] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *ICCV*, 2021.

[15] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything V2," in *NeurIPS*, 2024.

[16] M. Hu *et al.*, "Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, 2024.

[17] H. Lin *et al.*, "Depth anything 3: Recovering the visual space from any views," in *ICLR*, 2026.

[18] L. Piccinelli *et al.*, "UniDepthV2: Universal monocular metric depth estimation made simpler," *IEEE Tr. Pattern Anal. Mach. Intell.*, 2025.

[19] A. Bochkovskiy *et al.*, "Depth pro: Sharp monocular metric depth in less than a second," in *ICLR*, 2025.