AI Struggles to Recognize the Size of Landmarks

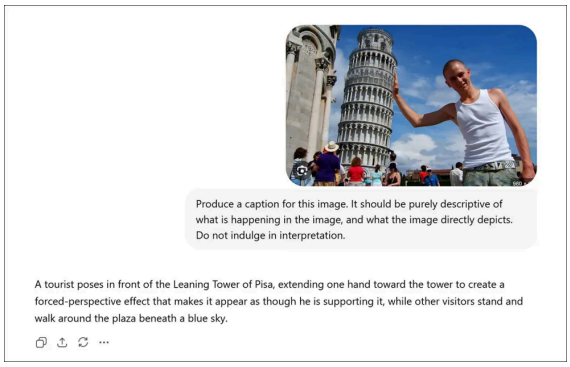
Originally published at https://www.unite.ai/ai-struggles-to-recognize-the-size-of-landmarks/

Vision Language Models understand monuments, but they still can't see the whole picture…

One of the earliest survival skills we develop is the ability to distinguish between things that are small or far away. We can blot out the moon with our thumb, without thinking that it is the size of a dime, because we have internalized an understanding of relative scale.

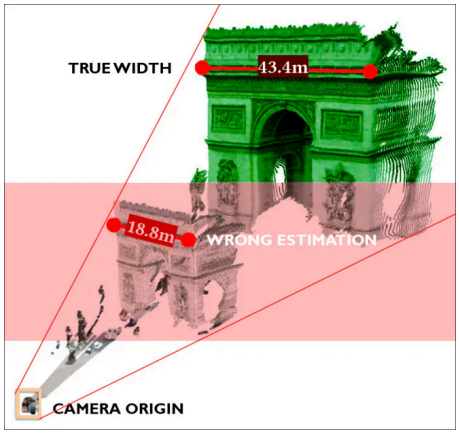
This is an unusually hard task for computer vision systems, since most of them rely on prior annotation, which does not help them to 'understand' scale in the same way as humans. To boot, beyond a certain and quite near limit, everything in the distance is beyond the ability of stereo vision to resolve – the car at the far end of the lot; the skyscraper in the distance beyond that; and the crescent moon rising over it…all are '2D' entities, for the majority of vision-based machine learning systems.

Of course, when a particular example of a 'distant' but misinterpreted object ends up well-represented in training data, systems that have seen this data can be tough to fool:



ChatGPT-5.5 is not remotely impressed with this classic tourist trope.

The less that a model's trained latent space contains such specific and oft-repeated information, the more it will need to be able to generalize and internalize the concepts of scale that we grasp at a young age. Without this, even famous examples can still cause misestimations of scale:



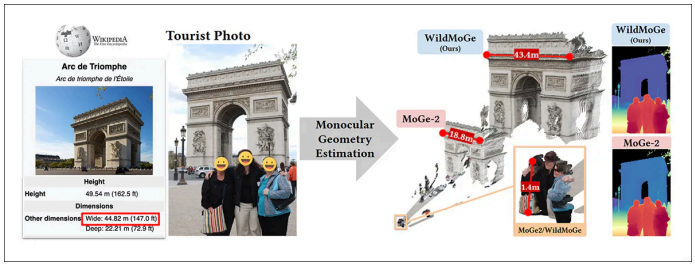
In this speculative example, adapted from the new paper that we are examining today, the camera's POV features the Arc De Triomphe in the background – but the system does not know what size it is, and makes an incorrect guess. Source

The danger, with specific and highly characteristic objects such as the Eiffel Tower, is that the system will resort to a shortcut of size estimation that is correct for the original model, but not correct for the multiple imitations of the Paris landmark that are equally beyond stereo resolution distance, yet are nowhere near as big.

Therefore it's important that vision systems approach novel (unseen) views with a ready skillset, and not just a bunch of 'cheat codes'.

Scaling Up

To this end, a new collaboration between the US and China offers a remediating dataset, together with an estimation method, that addresses the issue:



The new approach modifies a prior system through improved training material – data varied enough to provide a deeper understanding of depth issues.

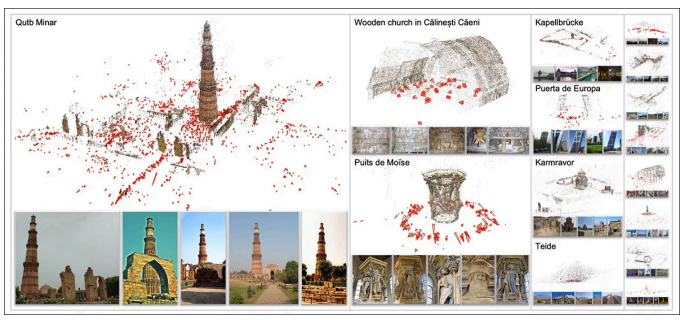
Launched together with an accompanying website, the *MetricScenes* initiative features data and code releases.

The paper states*:

> *'[We] found that current state-of-the-art methods frequently fail to estimate correct scene scale, leading to a persistent scale-collapse phenomenon in "in-the-wild" scenarios.*
>
> *'[The image above] shows an example where clear semantic references (people) are present, yet where models like MoGe-2 exhibit a significant scale inconsistency across the range of distances: the predicted metric scale for near-field objects is plausible – in this case, the tourists have a plausible height – yet the scale for far-field structures is dramatically underestimated – here, the Arc de Triomphe in the background is metrically predicted to be just 18.8 m wide, which is more than 2× smaller than the ground truth width (44.8 m).*
>
> *'MoGe-2 has posited a miniaturized landmark, despite cues to the contrary.'*

The Power of Three

The authors' new collection was assembled by combining three existing datasets: MegaScenes, AerialMegaDepth, and Stereo4D:



Example imagery from MegaScenes, which comprises part of the new curation. Source

The issue with the datasets that contribute to MetricScenes, when taken alone, is that they each apply to limited domains, such as POV car footage, or interior scenes, when a *combined* domain is needed in order to address the problem, and bring vision systems nearer to a human-style conceptual understanding of scale.

Each image is accompanied by RGB imagery, partially observed depth derived from Structure from Motion (SfM), Multi-View Stereo (MVS), or other geometric priors, together with a completed depth map generated through a new two-stage Poisson completion process, and associated camera metadata.

Fine-tuning the MoGe-2 framework on the new dataset *'significantly mitigates'* the scale collapse that the authors refer to, reportedly achieving superior results in open-domain scenes, and state-of-the-art performance on related benchmarks.

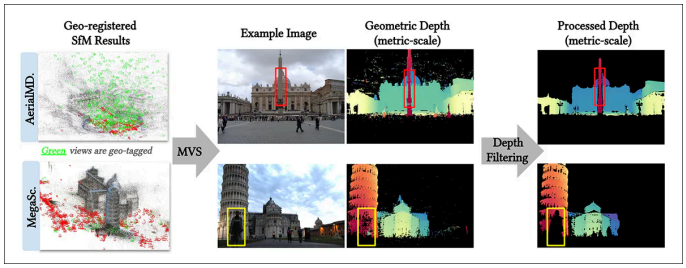
The new paper is titled *Honey, I Shrunk the Arc de Triomphe!*, and comes from four researchers across Cornell University and Shanghai Jiao Tong University.

Method

MetricScenes draws in part on the aforementioned *AerialMegaDepth* and *MegaScenes* – two collections of Internet photographs spanning historical archives, tourist images, and professional photography. Though MegaScenes offers large-scale Structure from Motion (SfM) reconstructions, these scenes lack any inherent real-world scale. To address this, geotagged imagery from online mapping services was used to align the reconstructions with known physical locations and dimensions.

Conversely, AerialMegaDepth already incorporates geotagged Google Earth views, providing metric-scale landmark reconstructions.

Potential reconstruction errors caused by visually similar but geographically-distant structures were addressed using MASt3R-SfM and the Doppelgangers++ classifier. After Multi-View Stereo (MVS) reconstruction, unstable depth estimates and depth-bleeding artifacts were filtered using a combination of stability checks and predictions from MoGe-2:



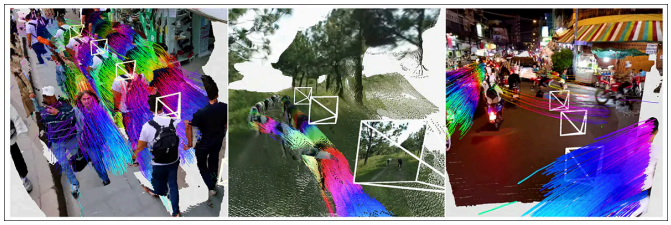
AerialMegaDepth derives real-world scale by combining Internet photographs with geotagged Google Earth views, while MegaScenes scenes are aligned to physical dimensions using geo-referenced street-level imagery. After Multi-View Stereo (MVS) reconstruction, unstable depth estimates and depth-bleeding artifac are filtered out, producing cleaner metric-scale depth maps suitable for training. Yellow boxes highlight transient objects removed during processing, while red boxes indicate corrected depth-bleeding regions.

Metric scale was then recovered through geo-referenced imagery. AerialMegaDepth already derives scale from Google Earth renderings captured from known locations, while MegaScenes was aligned to real-world dimensions using geotagged street-level imagery obtained from mapping services.

These images were matched to existing reconstructions with MASt3R, refined with the Doppelganger classifier, aligned with COLMAP, and scaled through RANSAC-based estimation using Earth-Centered, Earth-Fixed (ECEF) coordinates. Scenes with unreliable scale estimates, or poor registration quality, were discarded.

Seeing in Stereo

The MetricScenes collection also draws on the aforementioned Stereo4D dataset, which features thousands of real-world stereoscopic video sequences captured with VR180 cameras, offering a temporal dimension to the captures:

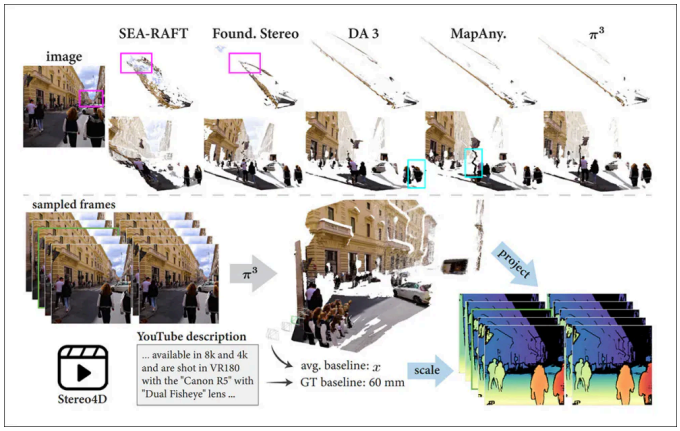


The Stereo4D dataset was built from stereoscopic Internet videos, combining camera poses, depth estimates, and motion trajectories to recover dynamic 3D scenes at scale. The resulting dataset contains hundreds of thousands of video clips represented as point clouds with long-range motion tracks, providing a large source of real-world 3D geometry and motion for training vision models. Source

Because the physical distance between the two camera lenses varies across different devices, only videos with documented camera configurations were used, allowing scene depth to be recovered at an accurate real-world scale.

Stereo4D originally relied on the optical-flow system SEA-RAFT to estimate scene geometry, but the authors found that imperfect camera calibration could distort reconstructed scenes, causing structures that should be parallel to converge unnaturally. Therefore, to improve accuracy, they replaced this approach with a multi-view reconstruction pipeline that jointly estimates camera poses and depth from multiple frames.

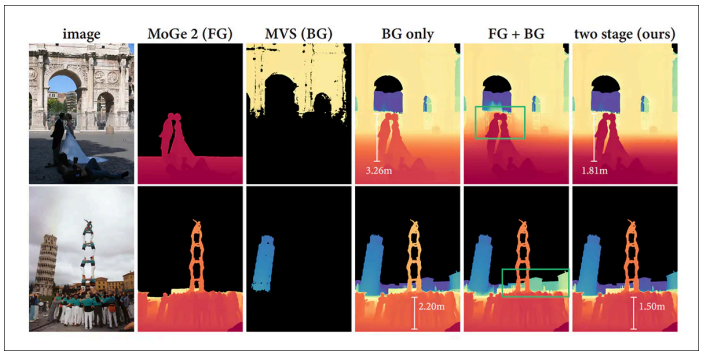
After comparing π^3, DepthAnything V3, and MapAnything, π^3 was selected for its geometric robustness and ability to preserve fine details:



Metric-scale depth recovery from Stereo4D. Standard stereo-matching methods can produce distorted geometry when camera calibration is imperfect, while π³ generates more consistent scene reconstructions, and preserves fine detail. The recovered geometry is then aligned to the known physical baseline of the stereo camera, yielding accurately-scaled metric-depth maps.

Because π³ reconstructs scenes at an arbitrary scale, the resulting depth maps were aligned to real-world dimensions using the known physical baseline of each stereo camera rig. Additional filtering removed low-quality frames, depth inconsistencies, calibration errors, and unreliable scale estimates.

Additionally, a two-stage depth-completion process was used, combining foreground predictions from MoGe-2 with background geometry from Multi-View Stereo (MVS), producing cleaner metric-scale training data with more consistent scale and sharper object boundaries:

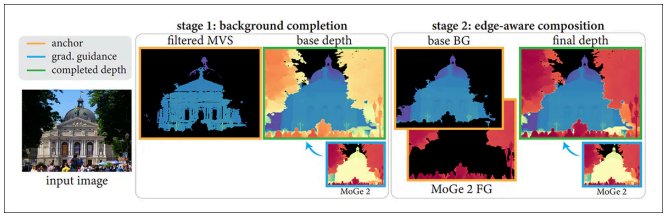


Two-stage depth completion. Using only background anchors can preserve scene structure while distorting overall scale, whereas combining foreground and background constraints in a single pass introduces scale drift and boundary artifacts. The two-stage approach maintains consistent metric scale across both near and distant objects while preserving clean object boundaries.

The authors observed that Internet photo collections often lack reliable foreground depth, while stereo imagery frequently misses distant background regions. Though MoGe-2 can infer dense geometry across an entire scene, its estimates tend towards the same *scale-collapse* problem that the project seeks to address. Therefore the two-stage depth-completion pipeline was designed to combine the strengths of MoGe-2 and Multi-View Stereo (MVS).

Background geometry was recovered using MVS-derived metric anchors, creating a base depth map with reliable large-scale structure. In a second stage, foreground estimates from MoGe-2 were reintroduced through an edge-aware completion process designed to preserve object boundaries while preventing scale drift and depth-bleeding artifacts.

The depth maps produced by this approach, the paper contends, were both visually complete and more consistent in real-world scale:



Two-stage depth completion pipeline. In the first stage, Multi-View Stereo (MVS) anchors are used to recover background geometry at a reliable metric scale. In the second stage, foreground estimates from MoGe-2 are reintroduced through an edge-aware composition process, producing a final depth map designed to preserve both large-scale accuracy and sharp local detail.

Data and Tests

The final MetricScenes collection comprises 47,579 exclusively real-world images covering 134 scenes from AerialMegaDepth; 29,583 images from 356 scenes from MegaScenes; and 22,549 frames taken from 1,725 videos from Stereo4D.

The collection, from which 10 scenes per source were held back as validation, covers exterior and interior contexts, as well as ground-level and aerial views, and urban as well as natural landscapes – a collated and cohesive context not available in any of the individual contributing collections.

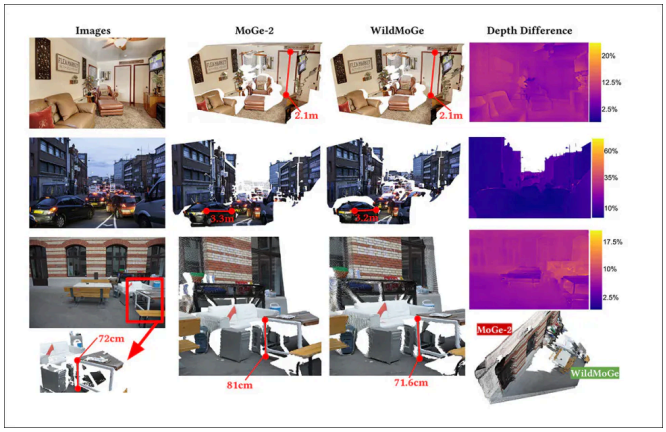
For an initial qualitative test, the authors fine-tuned the MoGe-2 *ViT-Large-Normal* model on the new MetricScenes dataset for 10,000 iterations at a batch size of 32 – effectively around three epochs. Cropping and general data augmentation approaches were taken from the original MoGe-2 tests, and training occurred at a learning rate of 1×10^{-6} (backbone) and 1×10^{-5} (all other parameters). For the qualitative test, depth reconstructions were undertaken by the fine-tuned *WildMoGe* model, pitted against base MoGe-2; DepthAnything V3; Metric3Dv2; UniDepth v2 ; and DepthPro:

Comparison of metric-scale landmark reconstruction. Ground-truth measurements from Google Maps are shown in the left column. Across unseen real-world landmarks, WildMoGe produces scale estimates more closely matching known dimensions, while MoGe-2, DepthAnything V3, and Metric3D V2 frequently underestimate the size of distant structures. UniDepth V2 often yields more plausible scales, but remains inconsistent, whereas DepthPro occasionally produces severe scale errors.

Of this result, the paper states:

> *'[WildMoGe] consistently recovers more accurate absolute scales across diverse landmarks, closely matching ground-truth dimensions (e.g., 31.4m vs. 32.4m for the Philadelphia Museum of Art, 46.7m vs 46.5m for Piazza della Signorina). MoGe-2, DepthAnything v3 and Metric3D v25 exhibit scale-collapse behavior, consistently underestimating the size of far-field structures.*
>
> *'UniDepth v2 produces more realistic scales but still deviates from ground truth, and DepthPro often fails to recover absolute scale, producing results that are orders of magnitude smaller than reality. Note that these scenes are absent from the training set.*
>
> *'This performance demonstrates that WildMoGe can generalize to unseen content, as opposed to simply memorizing training scenes.'*

To ensure the gains found were not limited to landmarks and large outdoor scenes, the authors also evaluated WildMoGe on ordinary indoor and street-level images, where it produced scale estimates broadly consistent with MoGe-2, while achieving better accuracy on an ETH3D courtyard scene:

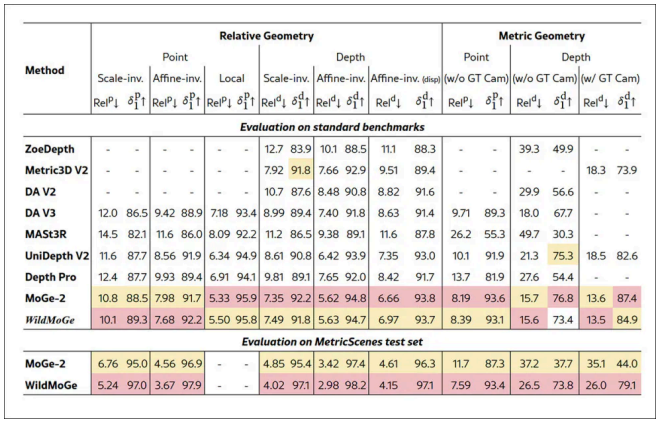


Comparison on standard scenes. Across ordinary indoor and street-level environments, WildMoGe produces scale estimates broadly consistent with MoGe-2, while achieving greater accuracy on the ETH3D courtyard benchmark, recovering object dimensions that more closely match ground-truth measurements.

To assess whether MetricScenes actually improved metric-scale reasoning, evaluation was performed both on a dedicated MetricScenes test set and on NYUv2; KITTI; ETH3D; iBims-1; GSO; Sintel; DDAD; DIODE; Spring; and HAMMER.

The authors note that obtaining dense ground-truth measurements for unconstrained Internet imagery remains difficult, meaning the MetricScenes labels are not perfect. Standard benchmarks were therefore included to verify that any gains did not come at the expense of general geometric performance.

Comparisons were made against MoGe-2; UniDepth V2; DepthPro; MASt3R; Depth Anything V2; Depth Anything V3; ZoeDepth; and Metric3D V2:



Method	Relative Geometry												Metric Geometry					
	Point						Depth						Point		Depth			
	Scale-inv.		Affine-inv.		Local		Scale-inv.		Affine-inv.		Affine-inv. (disp)		(w/o GT Cam)		(w/o GT Cam)		(w/ GT Cam)	
	$Rel^p\downarrow$	$\delta_1^p\uparrow$	$Rel^p\downarrow$	$\delta_1^p\uparrow$	$Rel^p\downarrow$	$\delta_1^p\uparrow$	$Rel^d\downarrow$	$\delta_1^d\uparrow$	$Rel^d\downarrow$	$\delta_1^d\uparrow$	$Rel^d\downarrow$	$\delta_1^d\uparrow$	$Rel^p\downarrow$	$\delta_1^p\uparrow$	$Rel^d\downarrow$	$\delta_1^d\uparrow$	$Rel^d\downarrow$	$\delta_1^d\uparrow$
Evaluation on standard benchmarks																		
ZoeDepth	-	-	-	-	-	-	12.7	83.9	10.1	88.5	11.1	88.3	-	-	39.3	49.9	-	-
Metric3D V2	-	-	-	-	-	-	7.92	91.8	7.66	92.9	9.51	89.4	-	-	-	-	18.3	73.9
DA V2	-	-	-	-	-	-	10.7	87.6	8.48	90.8	8.82	91.6	-	-	29.9	56.6	-	-
DA V3	12.0	86.5	9.42	88.9	7.18	93.4	8.99	89.4	7.40	91.8	8.63	91.4	9.71	89.3	18.0	67.7	-	-
MASt3R	14.5	82.1	11.6	86.0	8.09	92.2	11.2	86.5	9.38	89.1	11.6	87.8	26.2	55.3	49.7	30.3	-	-
UniDepth V2	11.6	87.7	8.56	91.9	6.34	94.9	8.61	90.8	6.42	93.9	7.35	93.0	10.1	91.9	21.3	75.3	18.5	82.6
Depth Pro	12.4	87.7	9.93	89.4	6.91	94.1	9.81	89.1	7.65	92.0	8.42	91.7	13.7	81.9	27.6	54.4	-	-
MoGe-2	10.8	88.5	7.98	91.7	5.33	95.9	7.35	92.2	5.62	94.8	6.66	93.8	8.19	93.6	15.7	76.8	13.6	87.4
WildMoGe	10.1	89.3	7.68	92.2	5.50	95.8	7.49	91.8	5.63	94.7	6.97	93.7	8.39	93.1	15.6	73.4	13.5	84.9
Evaluation on MetricScenes test set																		
MoGe-2	6.76	95.0	4.56	96.9	-	-	4.85	95.4	3.42	97.4	4.61	96.3	11.7	87.3	37.2	37.7	35.1	44.0
WildMoGe	5.24	97.0	3.67	97.9	-	-	4.02	97.1	2.98	98.2	4.15	97.1	7.59	93.4	26.5	73.8	26.0	79.1

Quantitative evaluation of relative and metric geometry. On the MetricScenes test set, WildMoGe outperformed MoGe-2 across every reported metric while remaining broadly competitive with ZoeDepth, Metric3D V2, Depth Anything V2, Depth Anything V3, MASt3R, UniDepth V2, and DepthPro on established benchmarks, indicating that improved metric-scale estimation was achieved without sacrificing general geometric reconstruction quality.

WildMoGe substantially improved metric-scale prediction on MetricScenes, outperforming MoGe-2 across every reported metric and achieving stronger metric-geometry and metric-depth scores than MoGe-2, Depth Anything V3, Metric3D V2, UniDepth V2, and DepthPro.

Performance on NYUv2, KITTI, ETH3D, iBims-1, GSO, Sintel, DDAD, DIODE, Spring, and HAMMER remained broadly comparable to MoGe-2. The authors attribute these gains to MetricScenes' metric supervision, which apparently helps to reduce scale collapse while preserving general scene reconstruction performance.

Conclusion

The MetricScenes solution to 'scale collapse' comes across as somewhat of a Heath-Robinson affair, in the paper – a hail-Mary melding and distillation of multiple datasets, each of which has some valuable viewpoint to contribute. It seems a little like trying to determine the shape of an elephant by touch.

Perhaps the most valuable service the paper offers is in calling greater attention to the issue, which seems to require some kind of novel or adapted universal standard. However, since such an innovation would interrupt the reproducibility and consistency of current methodologies, it would have to be very convincing.

* *My conversion of the authors' inline citations to hyperlinks.*

First published Thursday 11th June, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More