SINGLE-IMAGE FOCAL LENGTH ESTIMATORS HAVE A TELEPHOTO BLIND SPOT, AND THEIR BENCHMARKS SHARE IT

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-image focal length estimators are evaluated almost entirely on wide-to-normal lenses. We audit the four benchmarks recent calibration methods report on: none contains a photograph beyond 243 mm, and on two a constant fitted to the test labels beats every published method. Beyond this range published methods fail. On our benchmark and on three third-party sets that share no images or labels with it, every method trained on rendered panorama crops places at most 2% of telephoto ($\geq$135 mm) photographs within 25% of their true focal length.

The failure follows the training range. Fine-tuning only AnyCalib's decoder on photographs spanning 8–1200 mm raises its telephoto accuracy from 0% to 40–48% on all four sets, whereas the same fine-tune on photographs below 110 mm collapses again beyond that limit, so rendering is not the cause. Covering the full range with one model costs accuracy at 20–70 mm.

To make the range measurable we release **FocalCommons-100K**, 100,000 Wikimedia Commons photographs spanning 8–1200 mm, and **FocalBench**, a 3,177-image benchmark, and we measure how often EXIF focal tags disagree with structure from motion. The label-only statistics behind our audit apply to any regression benchmark and expose such coverage gaps before a model is run.

1 INTRODUCTION

Recovering a photograph's focal length from the image alone underpins single-image 3D reconstruction, metric depth estimation and augmented reality, and the metadata that would supply it is routinely missing: stripped by social platforms, absent from film scans, lost in screenshots. Recent methods predict a dense geometric field and fit a camera model to it, and report median errors of a few degrees of field of view (Jin et al., 2023; Veicht et al., 2024; Tirado-Garín & Civera, 2025). Throughout, we express focal length as its 35 mm equivalent, which fixes the field of view independently of sensor size, and count a prediction as correct when it lies within 25% of the truth.

Those errors are measured on a narrow slice of the problem. The four splits that GeoCalib (Veicht et al., 2024) and AnyCalib (Tirado-Garín & Civera, 2025) evaluate on, and later work reports (Armeni et al., 2017; Wang et al., 2020; Li & Snavely, 2018; Sarlin et al., 2022), were not built for calibration. None contains a photograph beyond 243 mm, and on TartanAir and LaMAR, each effectively a single focal length, a constant fitted to the test labels beats every published method (Table 1). A method can lead this suite without being tested on the long lenses used for wildlife, sport and distant architecture.

Outside that slice, published methods fail (Fig. 1). Every method trained on rendered panorama crops places essentially no telephoto photograph within 25% of its true focal length, on our benchmark and on three third-party sets: SR-RAW's optical zoom sweeps, FiveK's DSLR photographs and InFlux's calibrated video frames (Sec. 5). The failure follows the training range. The panorama generator GeoCalib and AnyCalib share samples fields of view equivalent to about 12–87 mm, extended to 110 mm by AnyCalib's aspect-ratio recut. Fine-tuning only AnyCalib's decoder on photographs that cover the full range raises the share of telephoto photographs it places within 25% from none to 40–48% on all four sets, while the same fine-tune restricted to photographs below 110 mm collapses

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

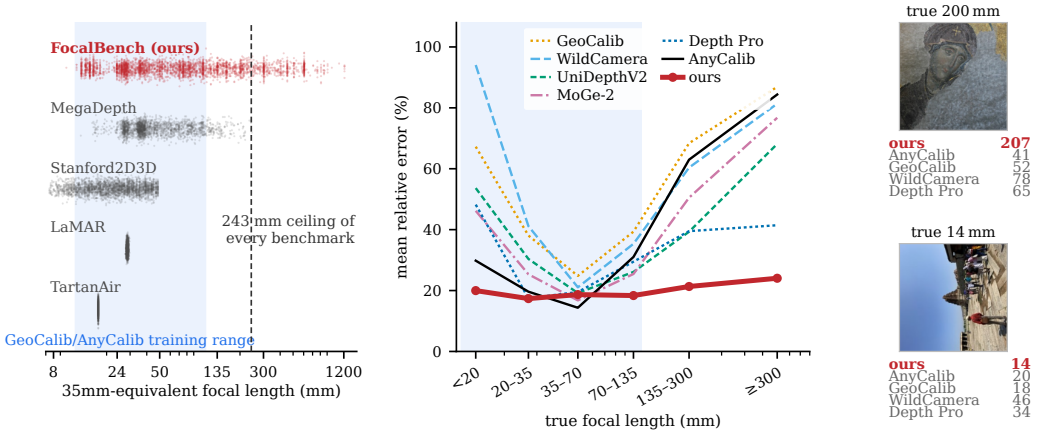


Figure 1: **Benchmarks sample a narrow focal band, and published methods fail outside it. Left:** every image of every benchmark; the four standard splits span 7–243 mm. Shaded is 11.7–109.7 mm: the panorama generator that GeoCalib and AnyCalib share reaches 86.8 mm, and AnyCalib's aspect-ratio recut extends it to 109.7. **Centre:** mean relative error by true focal length on FocalBench. **Right:** two benchmark photographs with the true focal length and each method's prediction, in mm.

again beyond it, although it never sees a render (Sec. 6). Covering the range is a trade: error rises in the 20–70 mm bands, where most photographs are taken.

The pattern is not specific to focal length: a regression benchmark whose labels span less than the range met in use cannot reveal an extrapolation failure, however well it ranks methods inside that range. Three statistics computed from the labels alone, the span, the effective number of distinct targets and the margin over a constant predictor, expose the gap before any model is run. And because a frozen backbone suffices once the labels cover the range, what the published models lack here appears to be label coverage rather than representation.

Our contributions are:

1. **An audit of the evaluation surface.** No standard calibration split tests focal lengths beyond 243 mm, two cannot distinguish estimating focal length from returning a constant, and the degree-based metric the field reports compresses telephoto error (Sec. 3).
2. **Evidence that the telephoto failure follows the training range.** The collapse replicates on three third-party sets. Fine-tuning AnyCalib's decoder on photographs covering the range repairs it on all of them, the same fine-tune restricted to AnyCalib's original range does not, and mixing panorama crops back in does not remove the mid-range cost (Secs. 5 and 6).
3. **FocalCommons-100K and FocalBench.** 100,000 photographs from Wikimedia Commons spanning 8–1200 mm, and a 3,177-image benchmark labelled by structure from motion or EXIF, every image redistributable as bytes (Sec. 4).
4. **A measurement of EXIF focal-tag disagreement** with structure from motion, the label source photo-collection training relies on: 12.4% of tags off by over 25%, with post-capture cropping a measured cause (Sec. 4.1).

We also train a plain DINOv2 regressor as a baseline. Sec. 7 reports how much of its FocalBench lead survives controls, and Sec. 8 where it does not lead.

2 Related Work

Single-image calibration. DeepFocal (Workman et al., 2015) regressed focal length from a photograph with a CNN trained on Flickr EXIF labels, the supervision FocaLens (Yan et al., 2017) and this work also use. Later methods regress horizon and field of view (Hold-Geoffroy et al., 2018;

Workman et al., 2016; Bogdan et al., 2018; López-Antequera et al., 2019), target in-the-wild imagery (Zhu et al., 2023) or images of people (Kocabas et al., 2021; Patel & Black, 2025), or reason from line segments (Lee et al., 2021). Recent methods predict a dense geometric field and solve for the camera: Perspective Fields (Jin et al., 2023) predicts per-pixel up-vectors and latitude, GeoCalib (Veicht et al., 2024) refines intrinsics with a differentiable optimiser, and AnyCalib (Tirado-Garín & Civera, 2025) fits them in closed form from ray directions. None evaluates beyond the wide-to-normal range.

Cameras from 3D foundation models. Monocular depth and point-map models increasingly output intrinsics or camera rays as a by-product, among them Depth Pro (Bochkovskii et al., 2025), UniDepth (Piccinelli et al., 2024; 2025b), MoGe (Wang et al., 2025b;c) and UniK3D (Piccinelli et al., 2025a). Multi-view reconstruction models such as VGGT (Wang et al., 2025a), MapAnything (Keetha et al., 2026) and Depth Anything 3 (Lin et al., 2026) predict cameras directly, and a focal length can be recovered from the point maps of DUSt3R (Wang et al., 2024) and π^3 (Wang et al., 2026). These models train on diverse 3D data, not panorama crops, and we find that they degrade far less outside the band (Sec. 5).

Narrow fields of view are ill-conditioned. As the field of view narrows, the ray bundle tends to a parallel projection and focal length becomes linearly dependent on camera distance. Photogrammetry puts the onset below 10°, "equivalent to a 200 mm lens on a 35 mm format" (Stamatopoulos et al., 2010; Stamatopoulos & Fraser, 2011); Strobl et al. (2009) plot calibration error against angular field of view, Beardsley & Murray (1992) made the point for vanishing points, and WildCamera notes it for learned estimation (Zhu et al., 2023). Those results concern recovering focal length jointly with exterior orientation, where the coupling is geometric and no dataset removes it. The failure we measure sets in earlier, where each method's training range ends, and it moves when the training data does (Sec. 6).

Datasets, benchmarks and dataset bias. Calibration models train mostly on panorama crops, which are exactly labelled but resolution-limited at the long end, and on structure-from-motion reconstructions of landmarks (Li & Snavely, 2018; Tung et al., 2024). A recent survey notes that the field lacks a unified benchmark (Liao et al., 2025). InFlux (Liang et al., 2025) and InFlux++ (Liang et al., 2026) calibrate zoom lenses in video, and Depth Pro reports focal accuracy on public photo collections including a zoom sweep (Zhang et al., 2019), without a breakdown by focal length. More broadly, benchmarks that sample a narrow slice of the input distribution overstate generalisation (Torralba & Efros, 2011; Recht et al., 2019), and models exploit whatever regularities their training data allows (Geirhos et al., 2020). Our audit is a label-coverage check in the first spirit, and the controls in Sec. 7 test for the second.

3 What the benchmarks can measure

We audit the four splits GeoCalib and AnyCalib both report on, using their own released labels (Table 1); ScanNet++ is access-gated, and the video and photo-collection suites above are not the splits calibration methods rank on. Two properties limit what the four can measure. The first is range: none extends beyond 243 mm, and only MegaDepth contains any photograph above 90 mm. The second is diversity: TartanAir and LaMAR each contain effectively one focal length. A constant fitted to their test labels beats every published method, so they measure accuracy at that focal length but cannot separate estimating it from returning a constant.

The standard metric compounds the problem. Error is reported in degrees of field of view, and degrees are not commensurable across the range: a 25% focal error is 9.5° of vertical field of view at 24 mm and 0.23° at 1200 mm, a factor of 41. A degree-based metric on wide-to-normal data is therefore doubly insensitive to telephoto error. We report relative focal error throughout and give degrees wherever we compare with published numbers.

Metrics and conventions. Focal lengths are 35 mm equivalents against the full-frame diagonal (43.267 mm). For a prediction $\hat{f}_{35}$, relative error is $|\hat{f}_{35} - f_{35}|/f_{35}$; we report its mean (**MRE**) and median (**MedRE**) in percent, and δ_τ, the share of images within $\tau\%$ of the truth. We call $\geq$135 mm *telephoto* and use six bands: below 20, 20–35, 35–70, 70–135, 135–300 and at or above 300 mm; a

benchmark	n	range (mm)	span	eff. focals	oracle constant	separation	≥90mm
MegaDepth	2,000	13–243	<u>18.7</u>×	5.4	30.2	+17.1	<u>12.9</u>
Stanford2D3D	2,415	7–48	6.5×	<u>5.6</u>	39.0	**+29.0**	0.0
TartanAir	2,000	17–17	1.0×	1.0	0.0	-11.4	0.0
LaMAR	2,000	28–30	1.1×	1.0	0.7	-3.0	0.0
FocalBench (ours)†	3,177	10–1200	**120.0**×	**11.0**	58.2	<u>+27.7</u>	**37.4**

Table 1: **What each benchmark can measure**, from its own labels. *span*: $f_{\max}/f_{\min}$; *eff. focals*: perplexity over half-stop bins; *oracle constant*: MRE (%) of the best constant focal, fitted on the test labels; *separation*: that minus the best published MRE.

band-balanced mean weights them equally. We also give median vertical-FoV error (°) and AUC@5 against published numbers. Grey cells are not ranked: § trained on the benchmark's images, ‡ its scenes in training. In tables, **bold** marks the best value and <u>underline</u> the second, and † marks our models and data.

4 FocalCommons-100K and FocalBench

Dataset. FocalCommons-100K holds 100,000 photographs from Wikimedia Commons, which accepts only freely licensed files (98.7% carry a recorded licence), 26.6% of them at or above 135 mm. Its provenance bounds the contribution: 87,847 (88%) come from MegaScenes (Tung et al., 2024), an internet collection of landmark reconstructions, re-labelled and re-balanced but not collected by us, and 12,153 are a new Commons harvest (Appendix C). The work is selection and labelling: MegaScenes is broad enough to support a telephoto split, and none had been built. Labels come from COLMAP (Schönberger & Frahm, 2016) bundle adjustment where a reconstruction exists and from EXIF otherwise.

Benchmark. FocalBench contains 3,177 photographs, 886 of them telephoto, in three slices defined by how the label was obtained: bundle adjustment, EXIF on landmark photographs, and EXIF on other photographs. The label source is the variable most likely to introduce a shared bias, and Sec. 7 uses the slices to test whether our conclusions depend on it. Every image ships as bytes with per-image attribution and licence. No image appears in both training and benchmark, which we verify exactly. The split is drawn by landmark, so no benchmark landmark has a training photograph, but scenes, the finer Commons categories that reconstructions are built from, can sit under neighbouring landmarks: of the 10,509 scene groups the benchmark draws on 766, and 585 of those also contribute training photographs. Sec. 7 measures what that overlap is worth by retraining without it.

4.1 EXIF focal tags disagree with structure from motion

EXIF supplies the labels for photo-collection training, ours included, and for 2,368 FocalBench images, so its reliability bounds what can be measured. On the 21,553 photographs carrying both an EXIF `FocalLengthIn35mmFilm` tag and a COLMAP focal length, we find 12.4% of tags off by over 25%, 77.3% of all disagreements in one direction (Table 8); Snavely (2008) had observed the effect on two landmarks. Below 70 mm, where most of these photographs lie, the main cause is post-capture cropping, which raises the true equivalent focal length while the tag still reports the lens. Images whose aspect ratio is evidence of a crop disagree by a median of 18.2% against 3.9% for standard ratios, and cropping accounts for at least 48% of the disagreement (Appendix A.2). At the long end the direction reverses and the cause is unresolved (Appendix A.1). Neither label is ground truth there: bundle adjustment loses its grip on focal length as the field of view narrows, and only three FocalBench photographs at or above 300 mm have an SfM label, so per-band error beyond 300 mm is measured against EXIF throughout.

method	all photographs			$\delta_{25}\uparrow$ by true focal length (mm)		
	MRE ↓	MedRE ↓	$\delta_{25}\uparrow$	<70	70–135	≥135
n		*3,177*		*1,671*	*620*	*886*
DeepCalib	53.5	51.9	24.1	45.9	0.0	0.0
Perspective Fields	49.3	46.5	29.0	53.3	4.8	0.0
WildCamera	50.8	45.7	29.3	43.6	30.2	1.9
GeoCalib	50.2	39.3	36.5	59.4	25.0	1.6
VGGT	42.3	33.0	39.8	43.7	55.6	21.2
UniDepthV2	36.4	27.9	46.5	57.0	55.8	20.3
Depth Pro	30.5	22.8	53.5	68.2	46.5	30.7
MoGe-2	36.7	27.1	47.5	65.8	54.5	8.2
AnyCalib	36.9	26.8	48.2	76.7	40.3	0.0
UniK3D	33.9	24.3	51.0	62.6	62.6	20.9
MapAnything	37.1	27.1	47.2	64.3	46.1	15.6
Depth Anything 3	39.4	25.6	49.3	64.0	55.3	17.2
AnyCalib + our data†‡	28.8 ± 2.0	14.9 ± 0.1	65.4 ± 0.6	73.0 ± 1.9	66.0 ± 5.9	50.7 ± 3.0
AnyCalib + our data, scene-disjoint†	33.6	16.4	65.0	77.2	61.8	44.2
MoGe-2 + our data†‡	30.6 ± 0.4	24.7 ± 0.8	50.7 ± 1.3	66.1 ± 0.6	53.7 ± 2.2	19.5 ± 6.3
Ours (ViT-L/14)†‡	19.6 ± 0.2	10.4 ± 0.2	74.9 ± 0.1	80.3 ± 1.0	76.3 ± 0.9	63.9 ± 1.2
Ours, scene-disjoint†	**27.9**	**13.1**	**69.7**	**79.1**	**68.2**	**52.9**

Table 2: **FocalBench.** MRE and MedRE are mean and median relative focal error (%); δ_{25} is the share of photographs within 25% of the true focal length (released checkpoint; ± one s.d. over three seeds). ‡: FocalBench's scenes are in training; grey rows are not ranked.

5 Published methods collapse outside their training range

On FocalBench. Table 2 and its per-band form, Table 7, are the central measurement. Every calibration-specific method roughly doubles its error from 35–70 to 70–135 mm and then fails. AnyCalib places **none** of the 886 images at or above 135 mm within 25% of the truth, and relaxing the threshold to 50% still admits only 8.1% of those 886 images. The collapse tracks training data more closely than method family. The three dense-field methods trained on rendered panorama crops, AnyCalib, GeoCalib and Perspective Fields, place 0.0%, 1.6% and 0.0% of telephoto images within 25%, and the two older panorama-trained methods, DeepCalib and CTRL-C (Appendix F.7), place none. The four trained on diverse 3D data, Depth Pro, VGGT, UniDepthV2 and MoGe-2, place 30.7%, 21.2%, 20.3% and 8.2%, and three newer 3D foundation models, UniK3D, Depth Anything 3 and MapAnything, place 20.9%, 17.2% and 15.6%. Across twelve methods this is a pattern rather than a controlled comparison; Sec. 6 supplies the intervention.

On MegaDepth's own labels. Split at 70 mm (Table 3), AnyCalib is within 25% on 94.2% of the 1,568 photographs below the line and 25.5% of the 432 above it, although MegaDepth stops at 243 mm. MegaDepth is not independent of every model: its eight test scenes all appear among our training landmarks, and seven of the twelve published methods train on MegaDepth itself (§ in the tables). AnyCalib trains on neither, so its collapse there stands, but improvements measured on MegaDepth do not isolate range; fine-tuning AnyCalib takes the long half to 60.0% within 25%, and we rely on the sets below instead.

On third-party photographs. To separate the collapse from our data, labels and landmarks, we score every method on three public sets with independent labels (Table 4; Appendix F.1): the 348 test photographs of SR-RAW (Zhang et al., 2019), each scene shot at up to seven optical zoom settings between 24 and 240 mm on a full-frame body; 4,624 uncropped DSLR photographs from FiveK (Bychkovsky et al., 2011), spanning 16–1,120 mm; and 1,470 video frames with calibrated intrinsics from the validation splits of InFlux and InFlux++ (Liang et al., 2025; 2026), spanning 22–335 mm. On all three, the panorama-trained methods place at most 2.0% of telephoto images within 25%, and the other published methods at most 16.0%, except Depth Pro on FiveK at 42.3%. The collapse belongs to the methods, not to our benchmark.

The boundary is in the training code. GeoCalib and AnyCalib render their training views from panoramas with a shared generator that samples vertical field of view uniformly on $[20°, 105°]$ at

method	δ_{25} <70 ↑	δ_{25} ≥70 ↑	MRE <70 ↓	MRE ≥70 ↓
AnyCalib	**94.2**	25.5	**8.7**	38.3
AnyCalib + our data†‡	91.5	60.0	11.2	28.3
GeoCalib	84.2	20.4	14.2	45.3
MoGe-2§	95.1	33.1	8.5	32.4
Depth Pro	93.9	**66.7**	10.2	**23.6**
UniDepthV2§	96.4	76.2	8.4	16.4
UniK3D§	97.6	89.4	8.2	12.5
MapAnything§	97.7	88.2	5.7	12.5
Depth Anything 3§	97.2	88.0	5.0	11.3
Ours (ViT-L/14)†‡	92.7	74.8	9.9	19.5

Table 3: **MegaDepth-2k, split at 70 mm** (1,568 photographs below, 432 at or above; third-party images and labels). δ_{25} and MRE (%) on each side. §: trained on this benchmark's images; ‡: its scenes are in training; grey cells are not ranked.

	SR-RAW (348)		FiveK (4,624)		InFlux (1,470)	
method	MRE ↓	δ_{25} ≥135 ↑	MRE ↓	δ_{25} ≥135 ↑	MRE ↓	δ_{25} ≥135 ↑
GeoCalib	36.7	2.0	33.7	0.6	43.6	0.0
VGGT	28.5	12.2	27.3	16.0	32.6	5.5
UniDepthV2	26.3	7.1	25.7	9.7	36.2	3.3
Depth Pro	24.2	10.2	**21.9**	42.3	29.9	15.5
MoGe-2	27.4	1.0	25.7	0.5	35.5	0.8
AnyCalib	26.8	0.0	28.2	0.0	32.4	0.0
UniK3D	24.1	12.2	23.3	8.5	33.4	8.3
MapAnything	34.2	4.1	28.1	11.1	42.4	6.1
Depth Anything 3	33.1	4.1	27.6	6.7	37.1	3.0
AnyCalib + our data†	**18.4**	**48.0**	26.7	44.8	26.6	**40.2**
Ours (ViT-L/14)†	18.9	22.4	23.5	**55.1**	**26.2**	32.4
Ours, scene-disjoint†	18.9	21.4	27.1	52.0	26.9	37.1

Table 4: **Third-party images.** MRE (%) over each set and δ_{25} at or above 135 mm; no image, label or pipeline is shared with our data. DeepCalib, Perspective Fields and WildCamera, omitted, place at most 1.1% at or above 135 mm on any set.

640×640,[1] which on a square frame is 11.7–86.8 mm and, because it is uniform in angle and not in $\log f$, dense at the wide end. AnyCalib re-cuts each sample to an aspect ratio drawn from $[0.5, 2]$,[2] which crops the long side and raises the equivalent focal by up to 26.5%, for an effective **11.7–109.7 mm**, median 29 mm. Each method fails past its own range. GeoCalib's error rises through 70–135 mm, beyond its 87 mm limit, while most of that band lies inside AnyCalib's range and its error there doubles but does not collapse. Beyond 110 mm, which neither has seen, AnyCalib's mean relative error is 63.0% at 135–300 mm and 84.4% above.

Range, not rendering. Rendered crops differ from photographs in more than focal range, so the boundary could reflect domain rather than range. WildCamera (Zhu et al., 2023), the one published calibration method trained on real in-the-wild photographs instead of renders, collapses as well: it places 1.9% of telephoto images within 25%, with 60.4% mean relative error at 135–300 mm and 81.3% above 300 mm. Rendering alone therefore does not explain the collapse. WildCamera's training focal range is not documented, so Sec. 6 settles the question with a controlled fine-tune.

[1] `siclib/datasets/create_dataset_from_pano.py`, byte-identical in both repositories. 640 is the render size; GeoCalib trains at 320×320 and AnyCalib at a constant 102,400 pixels.

[2] `siclib/configs/data/openpano-rays.yaml`.

6 COVERING THE RANGE IN TRAINING

A photo collection looks like the wrong training set for these methods. They predict a dense per-pixel field, of order 3×10^5 supervised values per image, and an EXIF tag supplies one number. For a pinhole camera with square pixels and a centred principal point, however, the bearing field $\mathbf{r}(u, v) \propto ((u - c_x)/f,\ (v - c_y)/f,\ 1)$ depends on the intrinsics alone, not on pose, gravity or depth, so one f_{35} value and the image size determine all of it.

AnyCalib. We fine-tune AnyCalib's decoder and head on FocalCommons-100K with its backbone frozen and its architecture and loss unchanged, so the training data is the only variable. Its target is exactly the field above, the ray parameterisation also used for pose estimation (Zhang et al., 2024); round-tripping f_{35} through AnyCalib's solver across 8–1200 mm agrees to 3×10^{-16}, so the supervision is exact. We train at the per-image aspect ratio AnyCalib's inference path derives; square centre crops cost 6.4 points of MRE (Appendix F.12). No benchmark image is in the training sample. On FocalBench, fine-tuning raises within-25% from 48.2% to 65.4% and, on the 886 telephoto images, from **0.0% to 50.7%**, or to 44.2% without the scenes it shares with the benchmark, the version Table 2 ranks; the released checkpoint's telephoto figure is 50.7 ± 3.0 and the aggregate 65.4 ± 0.6, one s.d. over three seeds. The gain is not learned EXIF noise: on the 809 structure-from-motion images mean error falls from 18.8% to 14.5%. The repair holds on the third-party sets, where the fine-tune gains 22 to 54 points of mean relative error at telephoto and follows InFlux's optical zoom with a within-video slope of 0.93, against the published model's 0.60 (Table 4). It is a trade. Mean relative error rises from 19.6% to 31.5% at 20–35 mm and from 14.3% to 30.3% at 35–70 mm (Table 7), and on Stanford2D3D, which is wide-to-normal, it costs 3.5 points; weighted by the focal distribution of MegaScenes, the fine-tuned model is worse than the published one, 30.1% against 23.9%.

Range, not domain. Two further fine-tunes of AnyCalib, two seeds each, isolate the cause (Appendix F.12). Trained on our photographs restricted to 11.7–110 mm, it collapses again beyond its range, placing 6.2% of images at 135–300 mm within 25% (0.0% on the second seed) and none at or above 300 mm, although it never sees a render; it recovers the middle of the range, with 20.2% and 19.9% mean relative error at 20–35 and 35–70 mm. Trained on half our photographs and half panorama crops drawn as AnyCalib's generator draws them, it keeps the telephoto repair but not the middle: at 20–70 mm it ties the full-range fine-tune, $+0.5$, $[-1.0, +2.0]$, and on the second seed -0.7, $[-1.6, +0.1]$. The mid-range cost is the price of one model covering 12–1200 mm, not of losing renders.

MoGe-2. MoGe-2 (Wang et al., 2025c) recovers focal length by a least-squares fit of focal length and depth shift to its point map, outside the computation graph. Fixing f leaves the shift as the only free parameter, so we solve for it and supervise the shift-corrected point map as a bearing field; solving the shift at the true focal length instead lets it absorb the error, and the loss falls while the focal estimate gets worse (Appendix F.11). The fine-tune reaches **19.5%** of telephoto images within 25% against the published MoGe-2's 8.2%. With one s.d. over three seeds the telephoto figure is 19.5 ± 6.3, so the direction is clear and the magnitude is not, and it loses accuracy on MegaDepth and Stanford2D3D (Table 5).

A plain regressor. As a baseline without a geometric head, we train a deliberately ordinary model: DINOv2 (Oquab et al., 2024) ViT-L/14, or ViT-B/14, with its last 16 of 24 (8 of 12) blocks unfrozen, mean-pooled into a two-layer MLP that regresses $\log f_{35}$ under an L_1 loss. Images are letterboxed to 252×252 rather than cropped, since cropping changes the quantity being predicted; the schedule and ablations are in Appendix F.19, and per-band scaling with training-set size in Appendix F.17. Its scene-disjoint version, the one FocalBench can rank, is best in every column of Table 2; the full ViT-L leads at 135–300 mm on all three third-party sets, but overall it leads only on SR-RAW and InFlux: on FiveK, most of which lies at 20–70 mm, Depth Pro is better, 21.9% against 23.5% (Table 4). On InFlux, where the zoom changes and the content does not, its predictions follow the zoom with a within-video slope of 0.99, against 0.66 for Depth Pro. Sec. 7 measures how much of its FocalBench lead survives controls. With architecture, schedule and seed fixed, it gains 6.1 points from data alone between 29.7K and 89.7K training images, where the larger set is also flatter in $\log f$ (Appendix F.18).

	MegaDepth			Stanford2D3D		
method	MRE ↓	FoV med ↓	AUC@5 ↑	MRE ↓	FoV med ↓	AUC@5 ↑
WildCamera[§]	14.2	2.9	41.6	35.0	10.9	11.3
GeoCalib	20.9	4.5	30.1	<u>11.6</u>	3.3	38.1
UniDepthV2[§]	10.1	2.2	49.2	19.0	6.2	21.9
Depth Pro	**13.1**	**2.7**	**42.9**	47.9	12.1	15.2
MoGe-2[§]	13.7	3.0	39.6	12.0	4.7	25.9
AnyCalib	<u>15.1</u>	<u>3.1</u>	<u>39.0</u>	**10.0**	**2.5**	**44.7**
UniK3D[§]	9.1	2.2	50.8	12.5	4.0	33.1
MapAnything[§]	7.2	1.2	66.4	15.0	5.0	26.2
Depth Anything 3[§]	6.4	0.9	71.2	23.2	4.6	32.7
AnyCalib + our data[†‡]	14.9	2.5	47.2	13.5	<u>3.0</u>	<u>39.6</u>
MoGe-2 + our data[†§]	17.1	5.6	21.0	25.4	13.1	7.3
Ours (ViT-L/14)[†‡]	12.0	2.1	50.7	24.2	6.9	19.8

Table 5: **Third-party benchmarks**: the two GeoCalib splits that can rank methods (Table 1). Relative focal error (%), median FoV error (degrees) and AUC@5. §: trained on this benchmark's images; ‡: its scenes are in training; grey cells are not ranked.

evaluated on / model	n	ours ↓	best published ↓	which	margin ↑	95% CI
all of FocalBench	3,177	19.6	30.5	Depth Pro	+10.9	$[+9.9, +12.0]$
SfM labels only	809	11.5	18.6	UniK3D	+7.1	$[+5.7, +8.6]$
unseen landmarks	292	19.0	26.2	Depth Pro	+7.3	$[+4.4, +10.0]$
landmark-disjoint retrain	3,177	27.9	30.5	Depth Pro	+2.6	$[+0.9, +4.3]$
size-matched control	3,177	19.4	30.5	Depth Pro	+11.1	$[+10.1, +12.2]$
top photographer out, 135–300 mm	427	21.8	35.5	UniK3D	+13.7	$[+11.0, +16.3]$

Table 6: **Every check that could remove the margin** (MRE, %). The best published method is recomputed per row; margin is its MRE minus ours, with a paired bootstrap interval over images. Last row: the worst band of Table 25.

A range control for the regressor. Restricting the regressor's training and validation data to 25–110 mm, with architecture, schedule and seed fixed, lowers error on the 1,675 benchmark photographs in that range from 17.7% to 13.5%, a gain of 4.2 points with a paired bootstrap interval of $[+3.2, +5.1]$, while outside the range the same model scores 62.1% against the full-range model's 21.8% (Appendix F.9).

7 Threats to validity

The collapse is measured on published methods that never saw our data, holds on both label sources and replicates on three third-party sets. The threats below concern the benchmark's labels, its overlap with our training data, and the margin of our baseline. Table 6 gives that margin under the checks that work as slices: the margin narrows from 10.9 points to 7.1 on SfM labels and 7.3 on unseen landmarks, and the retraining check below narrows it further.

Labels. 2,368 of the 3,177 labels are EXIF tags, so scoring against them could reward a model for learning the tags' errors; the 809 structure-from-motion labels do not. On those, our lead is 7.1 rather than 10.9 points, and Depth Pro, second on the EXIF-labelled images, falls to tenth of 15 (Table 12), although the two slices also differ in focal composition. The weak point is the long end: 3 of the 403 images at or above 300 mm carry an SfM label. At 135–300 mm, where 67 do, our lead on the SfM-labelled images is $+2.5$, $[-2.2, +7.8]$, so our telephoto *margin* is not separable from the label source. The *collapse* survives either label: on those SfM-labelled images AnyCalib places 0.0% within 25%, and it replicates on the third-party sets, whose labels come from optical zoom settings, DSLR metadata and calibration targets.

Landmark familiarity. Training and benchmark share no image, and near-duplicates are rare: perceptual hashing against all 89,651 training photographs finds seven benchmark images with a near-duplicate, 0.22% of the set, and excluding them moves no method's mean relative error by more than 0.04 points (about a tenth of the seed spread). Landmarks are shared: of the 1,935 landmark photographs in FocalBench, all but 292 come from a scene that also contributes training photographs. Removing every such scene from training and retraining costs the ViT-L **8.5 points** of mean relative error, $[+7.2, +9.9]$, relative to a control trained on a random subset of the same size, which itself costs nothing, -0.2, $[-0.6, +0.2]$. The control rules out the number of removed images but not their content. **Much of our FocalBench margin is familiarity**: without it the reported ViT-L scores 27.9% against Depth Pro's 30.5%, a lead of 2.6 points, $[+0.9, +4.3]$ (Appendix F.4); the margin actually lost to familiarity is 8.3 points. Photographers recur too, for 81.6% of benchmark photographs, and resampling them widens that interval to $[-0.1, +5.2]$. On the 582 photographs with neither landmark nor photographer in training, our model leads Depth Pro by 5.5 points, $[+2.7, +8.0]$, and the scene-disjoint model trails it by 4.9 in mean relative error, $[-10.4, -0.1]$, but leads by 12.0 points in δ_{25}. On SR-RAW and InFlux, which share neither, the scene-disjoint model ties the full one. Most methods, ours included, start from DINOv2, pretrained without focal labels on web images that include landmarks; we do not mark that. Neither the collapse, measured on methods that never saw this data, nor the AnyCalib repair, which keeps 44.2 of its 50.7 telephoto points without the shared scenes, depends on familiarity.

Shortcuts. Our telephoto data is rich in wildlife and aircraft, so a model could score well by recognising the subject. A nearest-neighbour lookup in frozen DINOv2 features, returning the median focal length of the ten most similar training photographs, places 35.4% of telephoto images within 25%: more than any published method, but about half of our model's 63.9%. An oracle told each image's Wikimedia category places **0.0%** of images above 300 mm within 25% against our 59.6%, and an oracle told each image's focal band beats our model in five of six bands but not above 300 mm, a band spanning two octaves, 24.0% against 28.2% (Appendix F.10).

Rendered images. FocalBench excludes the 832 perspective crops we rendered from panoramas. With their labels in the diagonal convention, every published method scored on them does better on the renders than on FocalBench, by 3.0 to 30.5 points, and ours by 0.9, so the exclusion favours our model (Appendix F.5).

8 Limitations and Conclusion

We built the benchmark we lead on. Benchmark and training data share a source: no landmark crosses the split, but scenes and photographers do, and our per-image intervals are too narrow for that clustering. This is why the collapse and the repair are also measured on three third-party sets, and why the model comparison is repeated without shared scenes and photographers (Sec. 7). The two label sources are not independent: COLMAP (Schönberger & Frahm, 2016) initialises its focal length from EXIF where a tag exists.

Telephoto labels are EXIF labels: 400 of the 403 images at or above 300 mm are EXIF-labelled, and the direction of EXIF–SfM disagreement reverses at the long end. Our telephoto margin is therefore not separable from the label source. The collapse is not affected, since the published methods predict normal-lens values for these photographs whatever their exact label.

We do not lead the standard benchmarks (Table 5). MegaDepth cannot rank our models, whose training data holds its scenes; among methods trained on neither its scenes nor its images, Depth Pro is best. On Stanford2D3D we are ninth of twelve on all three metrics, on TartanAir third of four among methods not trained on it (Appendix F.3), and on FiveK Depth Pro leads us overall. Every method receives the same images, at most 512 px on the long edge.

Conclusion. Single-image focal length estimators have been measured inside a band their benchmarks cannot see out of. Published methods fail past the focal range they were trained on, on our benchmark and on three third-party sets, and training on photographs that cover the range moves that boundary at a cost in the middle. Regression benchmarks should report their label span, effective number of targets and margin over a constant predictor, and methods should report error per band of the target.

Reproducibility Statement

All code, checkpoints and data needed to reproduce the experiments will be released under open licences. The code covers building FocalCommons-100K and FocalBench from Wikimedia Commons and MegaScenes (harvesting, labelling, balancing and splitting), training the regressor and the AnyCalib and MoGe-2 fine-tunes, and the evaluation harness, with adapters for every published method and preparation scripts for SR-RAW, FiveK and InFlux. Checkpoints for our regressors and for every reported fine-tune, including the range-restricted and mixed controls, will be released on Hugging Face. FocalBench is released as images with labels and per-image attribution, and FocalCommons-100K as a manifest of Commons filenames with labels; the third-party sets are public. We also release the per-image predictions of every method on every benchmark, so each table can be recomputed without running a model. Training schedules and hyperparameters are in Sec. 6 and Appendix F.19, seed spreads in Appendix F.15, and the split protocol in Appendix D. Rebuilding the manifest from scratch requires a copy of MegaScenes, but every image it lists can be downloaded from Commons by filename.

Ethics Statement

The dataset is assembled from Wikimedia Commons under the licences the photographs carry. 98.7% of rows record a licence and all recorded licences are free; per-image attribution and terms travel with the files, and the full text of every non-CC licence ships alongside, as GFDL 1.2 requires. Two photographs whose share-alike terms require an author Commons does not record were excluded from the benchmark rather than scored and withheld.

The photographs are of public places and some contain identifiable people. The qualitative figures are face-filtered. The uniform random samples deliberately are not: their purpose is to show what the data is actually like, and filtering them would misrepresent it.

AI use statement

In this work we used generative AI tools, an LLM-based coding and writing assistant, extensively. In the categories for which the ICLR 2027 policy requires disclosure, the assistant implemented methods (the training, evaluation, data-curation and paper-build code); cleaned and reformatted dataset metadata; proposed, designed and gave feedback on experiments and controls, including several reported here; launched and monitored experiment runs under the authors' direction; proposed and refined hypotheses; and helped interpret results. In the categories for which disclosure is recommended, it edited software code, created figures and tables, drafted and edited prose, searched and summarised literature, and suggested structure, title and keywords. It was not used to generate synthetic data, and the paper makes no mathematical claims.

The authors directed the project, decided which experiments to run and report, and checked the reported results against the released per-image predictions. All tables are generated from committed result files, and the main numbers in the text are recomputed from those files on every build by `paper/verify_claims.py`. The authors take full responsibility for the final content of this work, including all text, claims and artefacts.

References

Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. Joint 2D-3D-Semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

Paul Beardsley and David Murray. Camera calibration using vanishing points. In *British Machine Vision Conference (BMVC)*, 1992.

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations (ICLR)*, 2025.

Oleksandr Bogdan, Viktor Eckstein, François Rameau, and Jean-Charles Bazin. DeepCalib: A deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production (CVMP)*, 2018.

Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, 2020.

Yannick Hold-Geoffroy, Kalyan Sunkavalli, Jonathan Eisenmann, Matthew Fisher, Emiliano Gambaretto, Sunil Hadap, and Jean-François Lalonde. A perceptual measure for deep single image camera calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F. Fouhey. Perspective fields for single image camera calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision (3DV)*, 2026.

Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Jinwoo Lee, Hyunsung Go, Hyunjoon Lee, Sunghyun Cho, Minhyuk Sung, and Junho Kim. CTRL-C: Camera calibration transformer with line-classification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Erich Liang, Roma Bhattacharjee, Sreemanti Dey, Rafael Moschopoulos, Caitlin Wang, Michel Liao, Grace Tan, Andrew Wang, Karhan Kayan, Stamatis Alexandropoulos, and Jia Deng. InFlux: A benchmark for self-calibration of dynamic intrinsics of video cameras. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025.

Erich Liang, Caleb Kha-Uong, Chinmaya Saran, Sreemanti Dey, David W. Liu, Junhan Ouyang, Benjamin Zhou, and Jia Deng. InFlux++: Real and synthetic data for estimating dynamic camera intrinsics. *arXiv preprint arXiv:2607.05389*, 2026.

Kang Liao, Lang Nie, Shujuan Huang, Chunyu Lin, Jing Zhang, Yao Zhao, Moncef Gabbouj, and Dacheng Tao. Deep learning for camera calibration and beyond: A survey. *arXiv preprint arXiv:2303.10559*, 2025.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026.

Manuel López-Antequera, Roger Marí, Pau Gargallo, Yubin Kuang, Javier Gonzalez-Jimenez, and Gloria Haro. Deep single image camera calibration with radial distortion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Priyanka Patel and Michael J. Black. CameraHMR: Aligning people with perspective. In *International Conference on 3D Vision (3DV)*, 2025.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniK3D: Universal camera monocular 3D estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025b.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning (ICML)*, 2019.

Paul-Edouard Sarlin, Mihai Dusmanu, Johannes L. Schönberger, Pablo Speciale, Lukas Gruber, Viktor Larsson, Ondrej Miksik, and Marc Pollefeys. LaMAR: Benchmarking localization and mapping for augmented reality. In *European Conference on Computer Vision (ECCV)*, 2022.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Keith N. Snavely. *Scene Reconstruction and Visualization from Internet Photo Collections*. PhD thesis, University of Washington, 2008.

Clive Stamatopoulos and Clive S. Fraser. Calibration of long focal length cameras in close range photogrammetry. *The Photogrammetric Record*, 26(135):339–360, 2011.

Clive Stamatopoulos, Clive S. Fraser, and Simon Cronk. On the self-calibration of long focal length lenses. In *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume XXXVIII, Part 5, pp. 568–573, 2010.

Klaus H. Strobl, Wolfgang Sepp, and Gerd Hirzinger. On the issue of camera calibration with narrow angular field of view. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2009.

Javier Tirado-Garín and Javier Civera. AnyCalib: On-manifold learning for model-agnostic single-view camera calibration. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. MegaScenes: Scene-level view synthesis at scale. In *European Conference on Computer Vision (ECCV)*, 2024.

Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. GeoCalib: Learning single-image calibration with geometric optimization. In *European Conference on Computer Vision (ECCV)*, 2024.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A dataset to push the limits of visual SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Scott Workman, Connor Greenwell, Menghua Zhai, Ryan Baltenberger, and Nathan Jacobs. DeepFocal: A method for direct focal length estimation. In *IEEE International Conference on Image Processing (ICIP)*, 2015.

Scott Workman, Menghua Zhai, and Nathan Jacobs. Horizon lines in the wild. In *British Machine Vision Conference (BMVC)*, 2016.

Han Yan, Yu Zhang, Shunli Zhang, Sicong Zhao, and Li Zhang. Focal length estimation guided with object distribution on FocaLens dataset. *Journal of Electronic Imaging*, 26(3):033018, 2017.

Jason Y. Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *International Conference on Learning Representations (ICLR)*, 2024.

Xuaner Cecilia Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: In-the-wild monocular camera calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

This supplement documents the dataset and benchmark in the detail a reader needs to decide whether to trust them, and reports the analyses the main paper references. Everything here is regenerated from committed files by `paper/tables.py` and `paper/figures.py`.

A What the label disagreement is made of

A.1 Where the one-sidedness reverses

The one-sidedness is not uniform. It holds where the images are, and reverses where they are not. Among disagreements over 25%, the tag is the larger of the two in 10.6% of cases below 35 mm, 37.3% at 70–135 mm and 85.7% above 300 mm. Cropping cannot produce that direction, and neither can structure-from-motion noise as we can measure it: the high-confidence subset shows *more* reversal, not less. Digital zoom folded into the tag, a superzoom body reporting a nominal maximum, and bundle adjustment losing focal constraint as parallax shrinks would all produce it, and nothing we have separates them. We report it because it bounds the earlier claim rather than supporting it: below 70 mm, where four fifths of these photographs sit, the crop account survives; at the long end the label disagreement is real, sizeable, and unattributed on both sides.

A.2 How much of the label noise is cropping

How much of it is cropping. Enough to put a floor on it. The reference should be each camera's *own* native shape rather than a list of standard ratios: a 4:3 camera whose photograph arrives at 3:2 has been cropped, and a global list calls it standard. Taking the modal aspect ratio of every model with at least 20 photographs here (244 cameras covering 15,954 of the 21,553 pairs), those still matching their camera disagree by more than 25% at a rate of 6.2%, and those whose shape changed at 27.8%. The excess is **48% of the overall rate**, stable to two points across thresholds from 10 to 80 photographs per camera. It is a floor rather than an estimate: a crop that preserves the aspect ratio is invisible to any such test and sits in the matched group, alongside sensor sizes reported wrongly and lenses the body cannot read.

Taking the SfM label as correct, a method reporting below roughly 4.5% median relative error against EXIF labels, the median disagreement between the two sources, is measuring label disagreement rather than its own error.

B Additional tables and figures

Floats the body refers to.

Table 7 is the per-band form of Table 2, in mean relative error.

Table 8 is the EXIF label-noise measurement. Its three headline figures (12.4% of tags wrong by more than 25%, 77.3% of all disagreements one-sided, and the 18.2% against 3.9% split by crop evidence) are stated in Sec. 4.1.

C The dataset

Table 9 gives the composition. FocalCommons-100K is 100,000 openly-licensed photographs spanning 8–1200 mm, assembled from two sources. 87,847 come from MegaScenes (Tung et al., 2024):

	<20 mm	20-35 mm	35-70 mm	70-135 mm	135-300 mm	>=300 mm
n	*405*	*636*	*630*	*620*	*483*	*403*
DeepCalib	74.6	36.4	18.3	51.5	76.6	90.0
Perspective Fields	36.4	31.6	28.7	50.6	75.2	88.9
WildCamera	94.1	41.2	21.1	35.3	60.4	81.3
GeoCalib	67.2	38.1	24.8	39.3	68.4	86.9
VGGT	81.9	42.8	23.9	26.5	38.3	59.4
UniDepthV2	53.6	30.4	19.2	26.1	39.3	68.1
Depth Pro	48.2	**17.5**	19.5	29.5	39.5	41.4
MoGe-2	46.2	25.3	16.8	25.4	50.5	76.7
AnyCalib	29.8	19.6	**14.3**	31.0	63.0	84.4
UniK3D	45.1	29.7	20.4	**23.7**	37.2	62.1
MapAnything	48.5	25.6	20.5	31.0	44.6	69.8
Depth Anything 3	55.6	31.8	24.3	28.0	44.0	71.0
AnyCalib + our data†‡	19.9	31.5	30.3	26.8	32.9	29.1
AnyCalib + our data, scene-disjoint†	**17.4**	36.9	30.7	38.0	45.3	**28.8**
MoGe-2 + our data†‡	16.0	24.6	26.1	27.8	37.4	57.9
Ours (ViT-L/14)†‡	20.0	17.4	18.7	18.4	21.3	24.0
Ours, scene-disjoint†	21.7	32.0	25.9	28.1	**28.9**	**29.3**

Table 7: **Mean relative error by focal length.** Methods in the order of Table 2; † marks ours, ‡ a model whose training data holds FocalBench's scenes (grey, not ranked). Every calibration-specific method degrades sharply above 135 mm. Depth Pro degrades far less, which is consistent with the collapse being a property of training data rather than of the task — though see Table 12, where Depth Pro's standing depends on the label source. Bold marks the best value in each band *and every value tied with it*, where tied means within two s.d. of that band's seed-to-seed variation (Table 27; s.d. 0.2–1.1 points across the six). A single bold cell is therefore a resolved win and several are not.

over 21,553 images with both labels	%
disagree by more than 10%	26.7
disagree by more than 25%	12.4
median absolute disagreement	4.5
disagreements with SfM *larger*	77.3
among those over 25%	78.3
median disagreement, cropped aspect ratio	18.2
median disagreement, standard aspect ratio	3.9

Table 8: **EXIF focal tags are wrong, and one-sidedly so.** Every row compares an EXIF `FocalLengthIn35mmFilm` tag against a COLMAP focal for the same photograph. The one-sidedness is the evidence that the tag is the wrong one rather than the reconstruction: a post-capture crop raises the true equivalent focal length while the tag keeps reporting the lens, and images whose aspect ratio shows a crop disagree 4.7 times as much. Disagreement is $|f_{\text{exif}} - f_{\text{sfm}}|/f_{\text{exif}}$, a fraction of the tag being audited; normalising by the SfM focal instead gives 10.6% over 25% and a median of 4.4%, and leaves the one-sidedness unchanged.

38,803 from its released core and a further 49,044 found by walking the full mirror, which is where most of the telephoto mass came from. 12,153 are a new Wikimedia Commons harvest, split between subject diversity and a targeted long-lens sweep.

Balance. The dataset is deliberately flatter in $\log f$ than its sources. Sampling uniformly from MegaScenes gives 49.5% of images between 20 and 35 mm and 2.5% above 300 mm; we take every image in the scarce bands and subsample the abundant middle, which yields 12.4% below 20 mm and 26.6% at or above 135 mm. Section 6 shows why this matters: accuracy in each band tracks how much of the training set that band holds.

The stated 8–1200 mm range is populated throughout rather than defined by its endpoints. Counting by octave from 8 mm, the seven doublings hold 5,781, 26,317, 20,380, 20,431, 14,393, 9,558 and 3,034 images, and 106 lie above 1,024 mm. The extremes are genuinely thin (1% of the dataset is

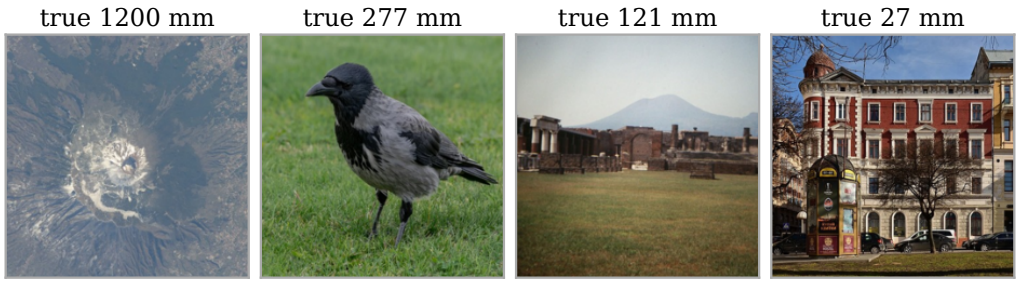


Figure 2: **What the collapse looks like.** One FocalBench photograph per focal band and what each method predicts. **These are selected**, by a stated rule: within each band, the image on which the median published method is furthest wrong and ours is within 15%. They show the failure mode the paper is about and are not a sample of typical performance: Table 2 carries the aggregate, Fig. 3 the eight-panel version, and Fig. 5 a uniform random sample in which nothing is chosen at all.

source	images	median (mm)	range (mm)	≥135mm (%)
MegaScenes (mirror walk)	49,044	72	8–1200	32.2
MegaScenes (v1 core)	38,803	40	8–1200	11.4
Commons, subject diversity	7,448	54	12–1200	24.2
Commons, long-lens harvest	4,616	400	8–1200	99.2
Commons, wide-angle harvest	89	14	12–19	0.0
total	**100,000**	**56**	**8–1200**	**26.6**

Table 9: **Where FocalCommons-100K comes from.** 87,847 images (88%) are MegaScenes, which we re-label and re-balance rather than collect; 12,153 are a new Wikimedia Commons harvest, of which the long-lens slice supplies most of the telephoto mass. The contribution is selection and labelling at scale, not collection.

below 13 mm and 1% above 750 mm), and a reader should treat the last octave as sparse rather than absent.

Labels. 35mm-equivalent focal length against the full-frame diagonal (43.267 mm), from COLMAP (Schönberger & Frahm, 2016) bundle adjustment where a reconstruction exists and the EXIF `FocalLengthIn35mmFilm` tag otherwise. Below 8 mm is fisheye, which breaks the pinhole assumption, and is excluded.

Licensing. 98.7% of dataset rows carry a recorded licence and all of those are free; the remainder are mirror rows absent from the index we join against, not images with unclear terms; Wikimedia Commons refuses non-free licences as policy. Per-image terms, author and source URL travel with the release, and the full text of every non-CC licence ships alongside, which GFDL 1.2 requires of anyone distributing a copy.

The benchmark is the part we redistribute as bytes, and the gate fails closed: every one of its 3,177 images ships. Two further photographs were excluded from the benchmark itself for the same reason (share-alike licences whose attribution condition we cannot discharge, because Commons records no author for them) rather than being scored on but withheld from the download; excluding them moves no figure in this paper by more than 0.03 points.

Reproducibility. The manifest is a released artefact rather than a derivable one: rebuilding it requires the local MegaScenes mirror. The *dataset* it describes is reproducible by anyone, because every image is re-fetchable from Commons by filename with the tool we ship.



Figure 3: **What the collapse looks like, two photographs per band.** The four-panel version is Fig. 2, where the selection rule and the caveat that goes with it are stated. The pattern is that published methods do not degrade gracefully past their training range: they return a normal-lens answer for a telephoto photograph, reading a 1200 mm satellite view of Vesuvius as 46 mm.

slice	images	label from	range (mm)	shippable (%)
commons diverse	1,242	camera EXIF tag	12–1200	100
megascenes exif	1,126	camera EXIF tag	10–1200	100
megascenes sfm	809	COLMAP bundle adjustment	13–845	100

Table 10: **FocalBench composition.** Three slices differing in how the label was obtained. The 809 bundle-adjustment labels are the control for the EXIF noise measured in Sec. 4.1, and the reason the benchmark could be larger but is not.

D THE BENCHMARK

Table 10 gives the composition of the three slices.

Splits. Of the 100,000 images, 6,700 are held out for the benchmark and 3,630 for validation; the remaining 89,670 train, of which 89,651 were fetchable at training time. FocalBench is drawn from them by capping each slice and band at 208 images and removing medium-confidence SfM labels, two images below 8 mm and two share-alike images without a recorded author; the remaining held-out images are unused rather than returned to training. The cap is why `commons_diverse` photographs, 7% of training, are 39% of the benchmark: the SfM pool holds only 67 images at 135–300 mm and 3 at or above 300 mm. A unit test checks that no image appears in both training and benchmark. Scenes overlap even though landmarks do not: 585 of the benchmark's 766 scene groups also contribute training photographs, and Sec. F.6 measures what that is worth.

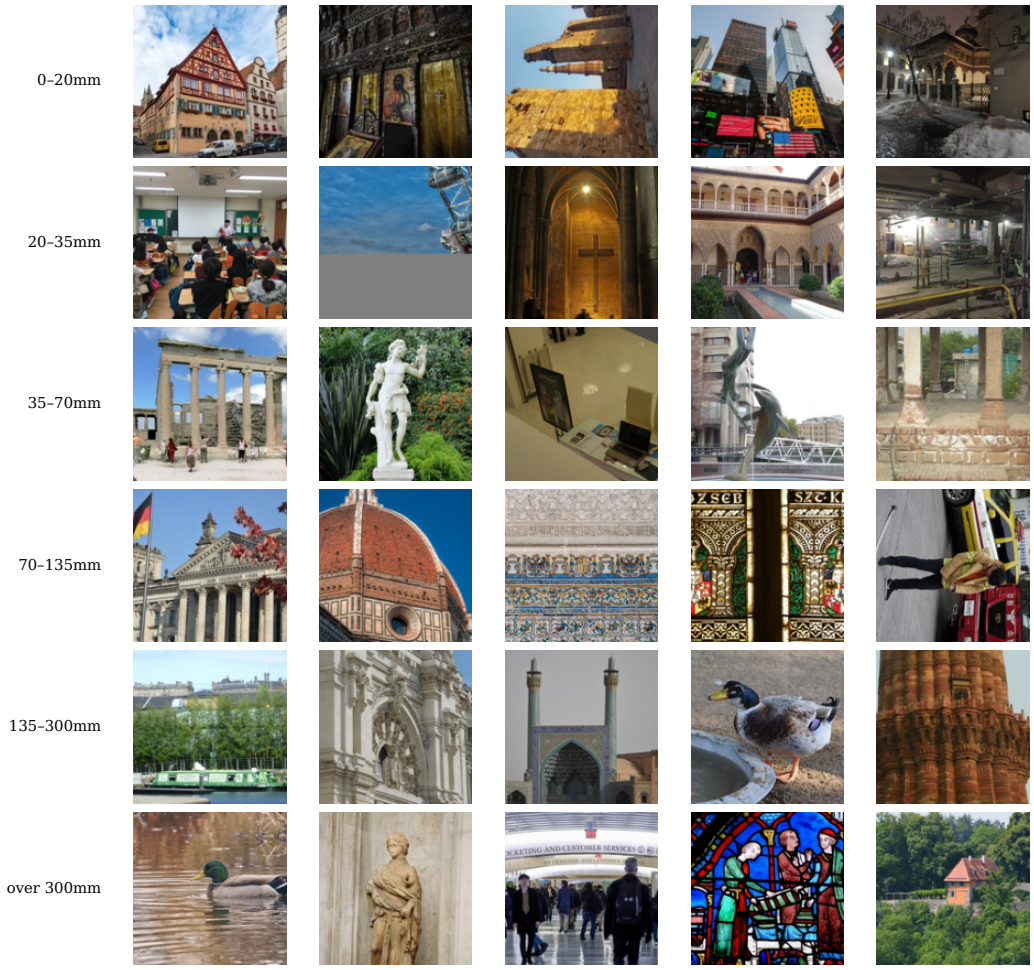


Figure 4: **A uniform random sample of the training set**, six images per focal band, fixed seed, nothing selected. Landmarks and interiors dominate the middle of the range, and the long end is what people actually point long lenses at: aircraft, wildlife, distant architecture and other people. Unlike Figs. 3, 7 and 8, this sample is not filtered for faces, so that it remains a uniform sample. Every image is freely licensed, with author and terms recorded in the manifest.

Geometry. 15.2% of benchmark images have a non-standard aspect ratio, which is evidence of a post-capture crop and the mechanism behind the one-sided EXIF error reported in Sec. 4.1.

E Additional comparisons

Figure 6 plots every method's prediction against the truth for every benchmark image, which is the collapse as a shape rather than a number. Figures 7 and 8 extend Fig. 3 in both directions: more cases where the published methods fail, and the cases where we do.

F Additional results

F.1 Third-party telephoto sets

Table 11 gives every method on the three third-party sets, with the audit statistics of Table 1 for each. All three separate methods from a constant predictor, by 20.6 to 26.2 points, unlike TartanAir and LaMAR. **SR-RAW:** the 50 test scenes, 348 uncropped full-frame photographs at 24–240 mm,

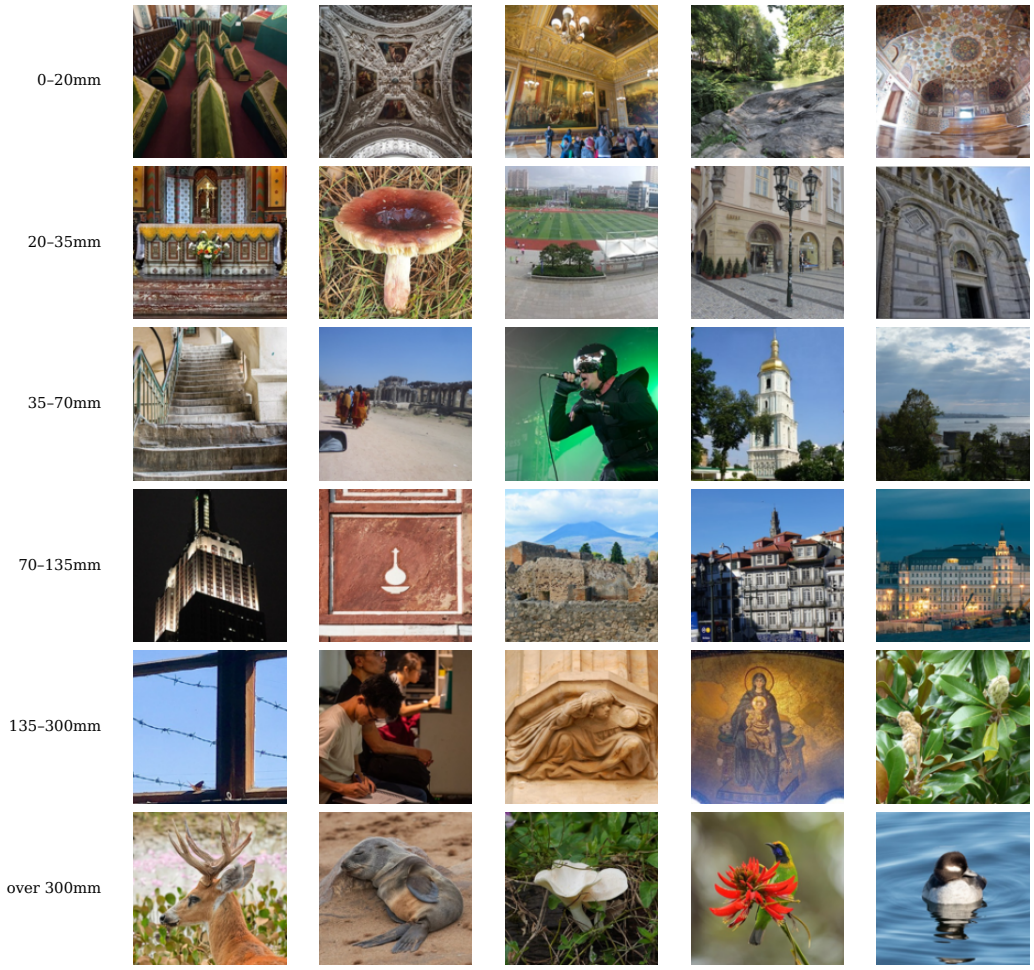


Figure 5: **A uniform random sample of FocalBench**, six per band, fixed seed, and the unselected counterpart to Fig. 3. Read it against that figure: these are the images our aggregate numbers are computed over, and they are ordinary photographs rather than the hard cases. Not filtered for faces, for the reason given in Fig. 4.

labelled by their EXIF 35 mm focal length; its training split was not available to us. **FiveK:** 4,624 photographs from 30 interchangeable-lens bodies, labelled by EXIF focal length times the body's crop factor, dropping frames whose aspect ratio differs from the sensor's by more than 3%, so its labels share the crop caveat of Sec. 4.1. **InFlux:** 1,470 frames from the validation splits of InFlux and InFlux++ with intrinsics from their calibration tables, at most five per video per band, scored as recorded with distortion and motion blur. Zoom slopes regress log prediction on log label within each scene or video. We lead at 135–300 mm on all three and overall on SR-RAW and InFlux; on FiveK, Depth Pro leads overall, by 1.6 points, $[+0.7, +2.3]$, and at or above 300 mm the two are level.

F.2 The leaderboard by label source

Table 12 scores every method twice: once on the 2,368 EXIF-labelled photographs and once on the 809 whose focal length comes from structure-from-motion and owes nothing to the camera. The ordering is stable at the top and not at the bottom, where Depth Pro falls from second to seventh, which Sec. 7 discusses.

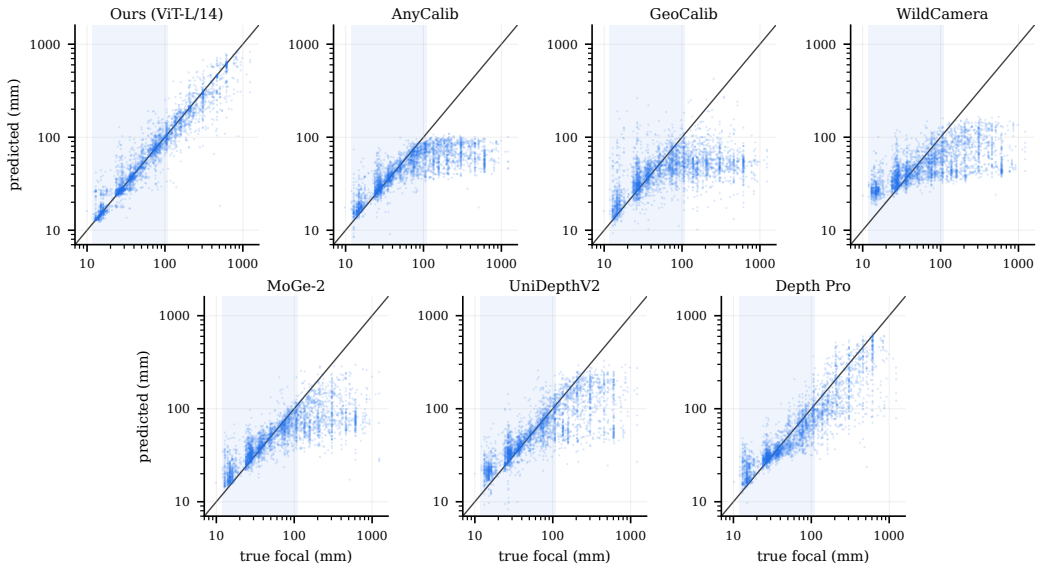


Figure 6: **Predicted against true focal length, every benchmark image, no selection.** The shaded band is 11.7–109.7 mm, the range the two panorama samplers reach between them. Inside it every method tracks the diagonal. Outside it the published clouds bend towards the horizontal: the prediction stops depending on the input, which is what the text means by returning a normal-lens answer for a telephoto photograph. AnyCalib's ceiling is the sharpest and sits almost exactly at the top of its sampling range. Depth Pro, trained on a different distribution, bends latest.

F.3 The splits we exclude, reported anyway

Table 5 ranks on MegaDepth and Stanford2D3D and omits TartanAir and LaMAR, because each holds one effective focal length and a constant predictor is therefore unbeatable on it: 0.0% on TartanAir, where every image shares a camera. Excluding them also removes **our worst result anywhere, TartanAir**: 39.2% against Depth Pro's 3.1. Table 13 gives both splits in full so a reader can apply the rule rather than take it. Our model does not exploit the single focal length these splits contain.

F.4 The margin under every check

Two analyses qualify the margin in Table 6.

Where the margin is, and what it is not. The 10.9 points are not spread evenly, and the strongest reading of this leaderboard is not the one we headline. Recomputing the best published method *within each band*, the same rule the table above applies across slices, we lead in four bands of six: $+9.8$ below 20 mm, $+5.3$ at 70–135 mm, $+15.9$ at 135–300 mm and $+17.4$ above 300 mm, against $+0.1$ at 20–35 mm, where Depth Pro is closest, and -4.4 at 35–70 mm, where AnyCalib leads. Two fifths of the benchmark lies in those two bands. Aggregating the per-band winners into an oracle that picks each band's best published method, a baseline that has to be told the focal length, which is the quantity being estimated, leaves our margin at $+6.1$ rather than $+10.9$, and the landmark-disjoint model's at -2.2. We headline the comparison against the best single published method because that is the model a reader can actually run, and we state the oracle because it is the strongest baseline this literature supports. The gap between the two is the claim: what we add is at the ends of the range, not in the middle, and for photographs between 20 and 70 mm AnyCalib is the better choice.

The margin without familiarity. Scored on all 3,177 images against Depth Pro, the model retrained with every FocalBench landmark removed still leads, by 2.6 points, $[+0.9, +4.3]$ by paired bootstrap over images. Resampling photographers instead of images widens the interval to $[-0.1, +5.2]$ (Sec. 7).

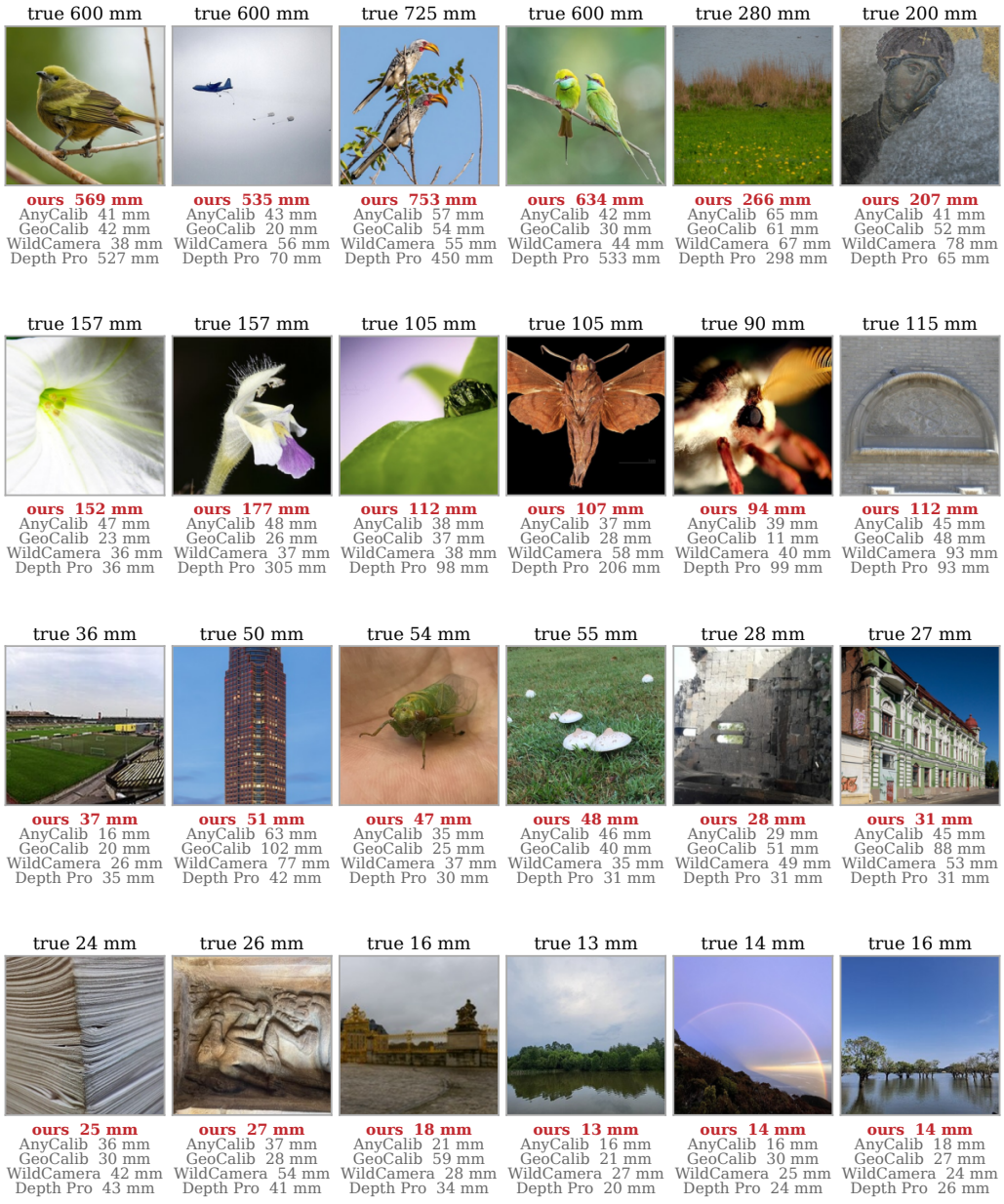


Figure 7: **Twenty-four further comparisons**, four per focal band, under the same rule as Fig. 3 and disjoint from it: the images on which the median published method is furthest wrong and ours is within 15%, after removing repeated subjects. Selected, and the whole range rather than the long end alone: the bottom row is 15–27 mm, where the published methods are supposed to be strong and where a 15 mm interior is read as 35–41 mm.

F.5 The rendered supplement

Table 14 scores every method on the 832 perspective crops we rendered from 120 panoramas, spanning 10.3–76.5 mm, beside FocalBench. Their labels were first computed against the 36 mm width and are now on the full-frame diagonal, like every other label here; the correction shortened them by a factor of 1.18; Sec. 7 states the comparison.



Figure 8: **Our eighteen worst predictions on FocalBench**, by relative error, chosen by the opposite rule to every other image figure here. Every one is an over-prediction, and most are close-focus photographs at short focal lengths where a shallow depth of field gives the background separation a long lens would: the model reads the blur as compression and answers two to four times too high. The published methods are closer on these images because they answer near 50 mm regardless.

Our rank on the standard splits. Counting only methods that did not train on a split's images or scenes, we are not ranked on MegaDepth, sixth of ten on LaMAR, ninth of twelve on Stanford2D3D and third of four on TartanAir. Focal range does not explain the low ranks: Stanford2D3D spans 7–48 mm, and on photographs in that band we score 17.6% against 19.6% overall. Rendering does not either, since our model does not lose on the rendered supplement once its labels are corrected. Two factors we have not separated remain: several methods ahead of us train on the split's source or on similar indoor and synthetic data, and our training set holds few indoor or synthetic images.

method	MRE ↓	MedRE ↓	δ_{25} ↑	δ_{25} <35	δ_{25} 35–70	δ_{25} 70–135	δ_{25} 135–300	δ_{25} ≥300	zoom slope
SR-RAW: 348 photographs, 24–240 mm, span 10.0×, 7.0 effective focals, oracle constant 50.3, separation +26.2, 28.2% at or above 135 mm									
DeepCalib	43.9	40.0	31.6	46.4	70.9	0.0	0.0	–	0.13
Perspective Fields	49.0	49.1	19.5	49.3	30.9	0.0	0.0	–	0.39
WildCamera	43.6	44.5	28.2	87.0	34.5	0.0	0.0	–	0.30
GeoCalib	36.7	32.8	40.5	92.8	58.2	15.5	2.0	–	0.35
VGGT	28.5	24.3	51.4	72.5	84.5	33.8	12.2	–	0.49
UniDepthV2	26.3	18.3	60.3	97.1	84.5	60.6	7.1	–	0.55
Depth Pro	24.2	18.2	60.3	91.3	85.5	60.6	10.2	–	0.67
MoGe-2	27.4	18.4	57.2	**98.6**	92.7	39.4	1.0	–	0.48
AnyCalib	26.8	18.3	60.1	**98.6**	**93.6**	53.5	0.0	–	0.51
UniK3D	24.1	18.2	60.1	92.8	86.4	53.5	12.2	–	0.62
MapAnything	34.2	30.5	44.5	95.7	66.4	16.9	4.1	–	0.47
Depth Anything 3	33.1	26.1	49.1	91.3	79.1	23.9	4.1	–	0.43
AnyCalib + our data†	**18.4**	**12.4**	**73.6**	97.1	80.9	74.6	**48.0**	–	0.82
MoGe-2 + our data†	28.9	22.5	54.9	60.9	88.2	67.6	4.1	–	0.71
Ours (ViT-L/14)†	18.9	13.4	71.0	92.8	86.4	**93.0**	22.4	–	0.75
Ours, scene-disjoint†	18.9	14.1	72.4	94.2	**93.6**	88.7	21.4	–	0.75
FiveK: 4,624 photographs, 16–1120 mm, span 70.0×, 7.5 effective focals, oracle constant 44.9, separation +22.9, 13.8% at or above 135 mm									
DeepCalib	40.3	39.0	33.1	28.8	75.3	0.0	0.0	0.0	–
Perspective Fields	40.1	37.1	33.0	50.5	47.0	12.4	0.0	0.6	–
WildCamera	34.3	29.1	43.3	70.5	56.5	17.5	1.5	0.0	–
GeoCalib	33.7	25.9	48.7	72.8	69.5	19.8	0.6	0.6	–
VGGT	27.3	22.1	55.4	59.7	76.8	43.6	19.6	5.9	–
UniDepthV2	25.7	18.3	60.6	78.7	82.4	38.7	13.2	0.0	–
Depth Pro	**21.9**	14.6	**70.6**	90.6	73.7	**59.1**	39.6	**49.7**	–
MoGe-2	25.7	17.9	60.5	**93.8**	81.6	26.8	0.6	0.0	–
AnyCalib	28.2	21.2	55.4	92.4	76.2	15.4	0.0	0.0	–
UniK3D	23.3	16.0	65.5	82.0	**85.6**	51.5	11.3	0.6	–
MapAnything	28.1	21.4	55.2	76.1	66.1	41.1	15.1	0.0	–
Depth Anything 3	27.6	20.5	57.4	81.3	72.3	38.0	8.9	0.6	–
AnyCalib + our data†	26.7	14.7	67.1	85.7	66.9	58.3	45.5	42.6	–
MoGe-2 + our data†	28.0	24.3	51.5	75.0	54.6	42.1	15.3	3.0	–
Ours (ViT-L/14)†	23.5	**14.0**	69.6	87.8	69.0	57.3	**57.2**	49.1	–
Ours, scene-disjoint†	27.1	14.7	66.8	87.1	66.4	52.1	53.2	48.5	–
InFlux: 1,470 video frames, 22–335 mm, span 15.3×, 7.8 effective focals, oracle constant 50.5, separation +20.6, 24.6% at or above 135 mm									
DeepCalib	47.9	50.2	22.6	42.2	54.7	0.0	0.0	0.0	0.10
Perspective Fields	54.9	59.5	13.6	40.9	17.8	1.2	0.0	0.0	0.37
WildCamera	45.1	48.9	25.9	69.7	36.4	5.3	1.3	0.0	0.25
GeoCalib	43.6	42.9	28.8	71.6	44.2	8.4	0.0	0.0	0.41
VGGT	32.6	30.0	43.5	62.2	**80.8**	30.2	6.6	0.0	0.42
UniDepthV2	36.2	35.0	36.7	**87.8**	51.4	14.4	3.6	0.0	0.43
Depth Pro	29.9	25.2	49.5	86.9	61.9	40.0	17.0	5.5	0.66
MoGe-2	35.5	34.5	37.4	82.8	58.3	16.7	1.0	0.0	0.47
AnyCalib	32.4	30.0	42.0	87.5	61.4	27.2	0.0	0.0	0.60
UniK3D	33.4	30.1	42.0	80.3	60.3	26.3	9.5	1.8	0.45
MapAnything	42.4	44.3	31.7	79.4	36.1	14.2	6.9	0.0	0.39
Depth Anything 3	37.1	35.2	36.7	82.8	52.2	17.7	3.6	0.0	0.46
AnyCalib + our data†	26.6	**18.7**	**60.8**	84.4	63.9	58.1	**44.3**	**16.4**	0.93
MoGe-2 + our data†	42.1	41.6	21.8	38.1	19.7	20.7	12.5	0.0	0.69
Ours (ViT-L/14)†	**26.2**	20.4	60.1	80.9	69.7	**59.8**	36.4	10.9	0.99
Ours, scene-disjoint†	26.9	21.2	57.8	78.8	62.2	55.6	42.0	10.9	1.04

Table 11: **Third-party images, in full.** Per set: the audit columns of Table 1, then every method's MRE, MedRE and δ_{25} (%), overall and by true focal length (bands under 20 photographs omitted). Zoom slope: within each scene or video with at least two focal lengths, the median slope of log prediction on log true focal length; 1 is a geometric estimator, 0 ignores focal length.

F.6 LANDMARK OVERLAP

Two measurements, and they disagree by a factor of three and a half.

The weak one compares a fixed model's score on benchmark images whose landmark it saw against those it did not: about two and a half points, taken inside the MegaScenes slices because elsewhere it degenerates: every `commons_diverse` image falls on one side and every rendered crop on the other. Several published methods score better on the unfamiliar half. Table 15 reports both halves, and the difference is not resolved once landmarks rather than images are resampled.

The direct one removes every landmark that appears in FocalBench from the training set and retrains. That costs the ViT-B 8.8 points of mean relative error, $[+7.2, +10.5]$ by paired bootstrap: three and a half times the weak estimate, and decisively resolved where the weak one is not. The effect replicates

method	EXIF labels (2,368)	SfM labels (809)
Ours (ViT-L/14)†	**22.4**	**11.5**
AnyCalib+ours†	33.6	<u>14.5</u>
UniK3D	39.1	18.6
AnyCalib	43.1	18.8
DA3	46.1	19.8
UniDepthV2	41.5	21.7
MoGe-2+ours†	33.4	22.5
MapAnything	41.9	23.0
MoGe-2	41.0	23.8
Depth Pro	<u>32.1</u>	26.0
GeoCalib	56.9	30.7
VGGT	44.9	34.6
Persp. Fields	54.1	35.1
DeepCalib	57.7	41.4
WildCamera	53.9	41.8

Table 12: **The leaderboard by label source.** The right-hand Our model leads both columns. **Depth Pro does not survive the change**, falling from two to ten of 15: its strong aggregate showing is in part an agreement with EXIF rather than with geometry, which weakens its use in Sec. 5 as the control separating task from training data. Only one of the three leaders is common to both columns. Every method scores better on the right; those images are easier as well as better labelled, so the columns are comparable within themselves and not to each other.

method	TartanAir	LaMAR
best constant predictor	0.0	0.7
WildCamera	43.4	5.5
GeoCalib	16.1	12.2
UniDepthV2§	7.0	6.4
Depth Pro§	3.1	15.4
MoGe-2§	12.1	5.9
AnyCalib	11.4	7.2
UniK3D§	2.9	12.7
MapAnything§	6.3	6.1
Depth Anything 3§	0.5	3.7
Ours (ViT-L/14)†	39.2	6.5

Table 13: **The two splits we exclude from the ranking** (MRE, %). TartanAir is synthetic, rendered with one pinhole camera, and LaMAR is a single capture rig, so each holds one effective focal length and a constant predictor is unbeatable; Table 5 omits both. **TartanAir is our worst result anywhere**. §: trained on this benchmark's images; ‡: its scenes are in training; grey cells are not ranked.

across backbones (Sec. 7 reports the ViT-L figure with a size-matched control), and Table 16 is the per-split evidence behind both. We rely on this measurement rather than the weak one.

F.7 What each method can emit

The ceiling argument needs no error metric. A prediction has to reach $0.75\,T$ to be within 25% of a target T, so the largest focal length a method ever produces fixes a ceiling above which no photograph is reachable, whatever that photograph contains. Table 17 is that ceiling for every method, computed from its own outputs with no labels and no error function involved. At the bottom, DeepCalib and CTRL-C cannot reach 48.8% and 48.5% of FocalBench; at the top our ViT-L reaches all of it. Fine-tuning AnyCalib on our data lifts its ceiling from 142 to 977 mm with its architecture untouched, which is the data result of Sec. 6 restated without an error metric.

method	FocalBench	supplement	Δ
Ours (ViT-L/14)†	**19.6**	18.7	-0.9
Depth Pro	30.5	24.8	-5.8
UniK3D	33.9	22.7	-11.2
UniDepthV2	36.4	28.8	-7.6
MoGe-2	36.7	21.8	-14.8
AnyCalib	36.9	**10.3**	-26.6
MapAnything	37.1	34.1	-3.0
DA3	39.4	33.6	-5.8
GeoCalib	50.2	19.7	-30.5

Table 14: **The rendered supplement.** FocalBench is 3,177 photographs; we also render 832 perspective crops from panoramas, and they are *not* part of it. Each of the eight published methods scored on it does better on the supplement than on all of FocalBench, by 3.0 to 30.5 points, and the two largest gaps belong to GeoCalib and AnyCalib, which train on rendered panoramas. Our ViT-L also does better on the supplement, by 0.9 points, and places two of nine on the supplement.

method	seen	unseen	Δ
Ours (ViT-L/14)†	**16.5**	**19.0**	+2.5
AnyCalib+ours†	22.9	25.8	+2.9
Depth Pro	27.9	26.2	-1.6
UniK3D	26.8	26.3	-0.5
MoGe-2	30.8	28.3	-2.6
AnyCalib	31.7	28.3	-3.4
MoGe-2+ours†	28.1	28.3	+0.2
UniDepthV2	29.1	28.8	-0.3
MapAnything	30.0	29.0	-1.0
DA3	30.7	29.2	-1.5
VGGT	35.5	34.0	-1.5
GeoCalib	40.8	40.5	-0.3
WildCamera	45.9	41.1	-4.9
Persp. Fields	45.7	43.4	-2.2
DeepCalib	49.6	48.8	-0.7

Table 15: **Landmark familiarity, measured.** Mean relative error on the 1,643 MegaScenes benchmark images whose landmark also appears in training, against the 292 whose landmark does not. No image is in both; these are different photographs of the same places. twelve of the 15 methods are *worse* on the familiar half, so familiarity is not simply a proxy for easier images. The three that gain most are the three trained on our data, which is the confound this table exists to size: it is worth $+2.5$ to our ViT-L and nothing to any published method. Across the three ViT-L seeds the premium is $+2.5$, $+2.7$, $+2.4$, so twice its s.d. is 0.4 points and ours is outside that.

CTRL-C, the clearest case. CTRL-C Lee et al. (2021) estimates camera parameters from an image and its line segments, trained on Google Street View and SUN360, both perspective crops rendered from panoramas. Table 18 reports it band by band: 66.3% within 25% below 20 mm, and then **zero** of the 1,504 photographs at or above 70 mm, across three consecutive bands. Its largest prediction anywhere is 49.7 mm.

CTRL-C targets upright architectural scenes and we run it unfiltered on general photography; inside the range it can express it does what it claims. The claim here is narrower and does not depend on the test set: a method trained on panorama crops inherits their focal range, and three consecutive bands at exactly zero is what that looks like from outside. The same test applies to our own models, in the last column of Table 17.

Why it is not a leaderboard row. Its line detector returns fewer than two usable segments on seven images, five of them one night scene, and it declines to answer there. Every row of Table 2 is a paired comparison over one shared set of 3,177 images; admitting a method covering 3,170 would

training set	FocalBench	MegaDepth	Stanford2D3D	TartanAir
ViT-L/14 — full training set (89,651)	19.6	**12.0**	24.2	39.2
benchmark landmarks removed (75,440)	27.9	12.1	27.3	45.7
random subset, same size (75,440)	**19.4**	13.0	**23.3**	**37.0**
ViT-B/14 — full training set (89,651)	**22.4**	13.9	33.7	69.2
benchmark landmarks removed (75,440)	31.4	**13.3**	39.0	69.3
random subset, same size (75,440)	22.6	13.8	**29.4**	**59.1**

Table 16: **What landmark familiarity is worth**, in mean relative error. Removing every landmark that appears in FocalBench costs 8.5 points *on FocalBench*. It matches the full model on MegaDepth, which is also landmark photography, so it has not stopped being a landmark calibrator — it has stopped recognising these particular places. It is worse than the size-matched control on the indoor and synthetic splits, which we do not claim to explain and report here rather than omit. Removing the landmarks also removes 15.9% of the training images, which is why the middle row has a size-matched control that keeps every landmark: that moves the score by at most 0.2 points, and in opposite directions on the two backbones (ViT-B/14 +0.2, ViT-L/14 -0.2), with an interval crossing zero. The measurement replicates across backbones — ViT-B/14 8.8, ViT-L/14 8.5 — which is why a number resting on one run of each is still worth stating: 16 of 24 blocks unfrozen for the ViT-L and 8 of 12 for the ViT-B, 25 epochs, seed 1.

method	largest prediction	reaches	unreachable %	99th pct.	unreachable %
DeepCalib	49	66	48.8	48	49.0
CTRL-C	50	66	48.5	38	54.0
Persp. Fields	96	129	28.4	77	34.6
AnyCalib	107	142	25.4	94	28.9
WildCamera	160	213	16.2	137	20.4
MoGe-2	294	392	8.1	207	13.3
UniDepthV2	320	426	7.0	236	9.9
UniK3D	358	478	5.4	282	8.5
MoGe-2+ours†	369	492	5.3	304	7.5
MapAnything	376	501	4.9	246	9.5
GeoCalib	412	550	4.4	134	20.8
VGGT	448	597	4.1	313	7.2
DA3	450	600	4.1	247	9.4
Depth Pro	639	852	0.4	525	1.3
AnyCalib+ours†	733	977	0.2	609	0.7
Ours (ViT-L/14)†	1166	1555	0.0	631	0.4

Table 17: **What each method can emit.** A prediction must reach $0.75\,T$ to be within 25% of a target T, so the largest focal length a method ever produces on FocalBench fixes a ceiling above which no image is reachable — whatever the photograph contains. Columns: the largest prediction made anywhere on the benchmark, the target above which it can no longer be within 25%, and the share of the benchmark that rules out; the last two repeat that with the 99th-percentile prediction, which one outlier cannot set. This is arithmetic on each method's own outputs; no labels and no error function enter it. DeepCalib cannot reach 48.8% of the benchmark and Ours (ViT-L/14)† reaches all of it, and the ordering is nearly the leaderboard's — which is the claim: at the bottom of that table the failure is not inaccuracy but a head that cannot represent the answer. Fine-tuning AnyCalib on our data lifts its ceiling from 142 to 977 mm and its unreachable share from 25.4% to 0.2%, with its architecture untouched — the same result as the leaderboard, in a form that does not involve a metric. Sorted by the last column and nothing is marked best: this is a range, not a score. † marks our models and data.

drop those seven images from every published row too, moving Depth Pro from 30.5 to 30.3 among other printed figures.

band	n	MRE ↓	δ_{25} ↑	δ_{50} ↑
<20 mm	401	22.9	66.3	89.5
20-35 mm	635	25.7	47.2	98.0
35-70 mm	630	49.9	8.3	43.2
70-135 mm	618	73.6	0.0	0.6
135-300 mm	483	87.3	0.0	0.0
>=300 mm	403	94.8	0.0	0.0

Table 18: **CTRL-C: an output range, not an error rate.** An ICCV 2021 calibration method trained on Google Street View and SUN360, both perspective crops rendered from panoramas. Its largest prediction anywhere on this benchmark is 49.7 mm, so it cannot land within 25% of any target above about 66 mm — 48.5% of FocalBench, ruled out before accuracy is in question. That is a fact about the network's output distribution and not about the photographs we chose: no test set reaches a focal length the head does not emit, which is why 70-135 mm, 135-300 mm, >=300 mm sit at exactly zero. Inside the range it can express it is a real method, placing 66.3% of the sub-20 mm photographs within 25%; the middle 90% of its predictions span 15 to 33 mm, 1.1 octaves. It is reported here rather than in Table 2 because its line detector returns fewer than two segments on seven images, and seating a method covering 3,170 in a paired table would drop those seven from all twelve published methods and move 43 of their reported figures, across 12 of the 12 published rows.

checkpoint	MRE ↓	δ_{25} ↑	<20	20-35	35-70	70-135	135-300	>=300
`pinhole` (reported)	**36.9**	**48.2**	29.8	19.6	**14.3**	31.0	63.0	84.4
`edit` (crop-augmented)	39.9	40.1	39.6	31.2	17.6	**28.7**	**59.3**	**82.5**
`gen`	37.0	47.6	**27.4**	**18.3**	15.5	32.3	63.7	84.8

Table 19: **All three released AnyCalib checkpoints.** The `edit` variant is trained with aspect-ratio crop augmentation, which extends its effective focal range past the generator's 86.8 mm, and it is the natural candidate for explaining the collapse away. It does not: telephoto error falls by 3.7 and 1.9 points and remains at 59% and 82%, while its aggregate is worse than the checkpoint we report. We report `pinhole` throughout because it is AnyCalib's best checkpoint on this benchmark, not its weakest.

F.8 AnyCalib's own checkpoints

AnyCalib ships four checkpoints, of which we score three; `dist` is left out. We report `pinhole`; `edit` is trained with aspect-ratio crop augmentation that extends its effective focal range and is the natural candidate for explaining the collapse away. Table 19 shows it does not.

F.9 The range expert and the cascade

Table 20 is the restricted-training-split experiment behind Sec. 6: identical images, architecture, schedule and seed, with the training and validation splits held to 25–110 mm.

Table 21 then gives the three rows the cascade is built from: a generalist, that same expert, and the cascade routing between them on the generalist's own estimate, against the two oracles that bound it.

F.10 What a shortcut could buy

Table 22 bounds the two shortcuts a reader might suspect behind our telephoto result: recognising the subject, and recognising the octave. Both are oracles, told the answer they would have to infer, so they bound rather than estimate.

scored on	n	generalist	expert	Δ
25–110 mm (the expert's range)	1,675	17.7	**13.5**	+4.2
outside it	1,502	**21.8**	62.1	-40.4
all of FocalBench	3,177	**19.6**	36.5	-16.9
median, 25–110 mm	1,675	8.7	**7.8**	+0.9

Table 20: **What a range expert buys, and what it costs.** Identical images, architecture, schedule and seed; only the *training split* is restricted to 25–110 mm. Inside that band the expert gains 4.2 points, paired bootstrap $[+3.2, +5.2]$; outside it it is unusable. The median row is the result rather than a detail: it moves by 0.9 points against 4.2 in the mean, so what covering $150\times$ of focal length costs is the rate of gross failure and not everyday precision. Bold compares the two models within a row; Δ is positive where the expert wins.

model	mean	<20	20-35	35-70	70-135	135-300	>=300
generalist	19.6	**20.0**	17.4	18.7	18.4	**21.3**	**24.0**
expert (25–110 mm)	36.5	80.3	**14.8**	**14.0**	**16.6**	48.7	77.9
cascade (image-routed)	**19.1**	21.2	15.1	15.3	18.2	24.0	24.4
oracle: true band	17.4	20.0	14.1	14.0	15.3	21.3	24.0
oracle: better model per image	15.2	19.3	11.6	11.4	11.1	19.6	23.9

Table 21: **Routing between a generalist and a range expert**, on real photographs. The expert is trained on 25–110 mm and is the best model we have inside that range and useless outside it, which is the capacity argument in three rows. The cascade and its limits are discussed in Sec. 6; its 0.54-point gain is one run, against a seed-to-seed s.d. of 0.21 for this backbone, so we do not claim it, and it is inconclusive besides against an oracle headroom of 4.39.

F.11 What the MoGe objective optimises

Table 23 is the measurement behind Sec. 6's negative result. It is worth a supplement section rather than a footnote because the failure is not specific to MoGe: any method that recovers intrinsics by fitting a model with nuisance parameters can be trained to satisfy a residual evaluated at the true answer while the fit itself moves the other way. The loss curve gives no warning: ours fell by a factor of eight.

F.12 Every fine-tune we ran

Table 24 lists all seven arms, so the reported one can be read against the alternatives rather than on its own. The recipe was fixed before the MoGe run so that both methods get the same one, rather than chosen on the benchmark; it happens to be the best of the seven on mean relative error and on δ_{25}, and second on telephoto δ_{25} behind the square-training-frame arm.

The two differ only in which manifest they sample from, and the difference is instructive because it is the same range trade Sec. 6 measures, appearing inside a published architecture. Equal-weight sampling asks for 2,000 images per band. The 40.6K manifest cannot supply that at the ends (it holds 1,231 images above 300 mm), so the realised sample is 11% long-lens. The 100K manifest can, so the realised sample is 17%. What follows is an exchange, and not a clean one, in δ_{25}: $+18.4$ points above 300 mm and $+8.6$ below 20 mm, against -5.7 at 35–70 mm and -2.7 at 20–35 mm. The ends prefer the flatter manifest and the middle prefers the other, which is why the aggregate, whose mass is in the middle, separates the two arms by less than a point. Reweighting the training sample is the variable that matters, and the telephoto-only arm is the control showing the cost comes from what was removed rather than from fine-tuning as such.

F.13 Provenance concentration

Table 25 recomputes the margin in every band with that band's largest contributor removed. Every interval in this paper is a per-image bootstrap and photographs by one person are not independent draws, so this sizes the exposure rather than correcting for it. The band we expected to be worst

	<20 mm	20-35 mm	35-70 mm	70-135 mm	135-300 mm	>=300 mm
category oracle	333.5	175.0	84.4	37.9	60.9	82.6
focal-band oracle	**7.9**	**7.6**	18.4	**16.2**	**18.9**	28.2
Depth Pro	48.2	17.5	19.5	29.5	39.5	41.4
AnyCalib	29.8	19.6	**14.3**	31.0	63.0	84.4
Ours (ViT-L/14)†	20.0	17.4	18.7	18.4	21.3	**24.0**

Table 22: **What a shortcut could buy** (MRE, %). The *category oracle* predicts the mean focal length of the image's own Wikimedia category, computed on training images: it is what a model that recognised the subject perfectly and measured nothing would score. The *focal-band oracle* is told which octave the image belongs to and answers that octave's best constant. Both use test-set information no model has. Knowing the subject is worth almost nothing — at or above 300 mm the category oracle places 0.0% of images within 25%, against our 59.6% — which answers the objection that our telephoto result is wildlife recognition. Knowing the band is worth a great deal in the middle of the range and beats us there, and cannot be had without already knowing the answer.

	residual at the true focal, deg		recovered focal error, % ↓	
	<135 mm	≥135 mm	<135 mm	≥135 mm
published MoGe-2	1.286	0.424	26.8	61.1
shift pinned at the label	0.203	0.053	29.8	69.8
shift solved jointly	2.977	0.941	49.4	42.8

Table 23: **The proxy runs backwards.** Rows are the published model and two diagnostic fine-tunes, neither of them the reported MoGe-2 + our data. Columns 1–2 are the angle between the shift-corrected point map and the field the true focal induces, with the shift solved *against that focal* — the objective the second row trains on, measured the same way for all three. Columns 3–4 are what the benchmark scores: the focal MoGe recovers with focal and shift solved jointly, as at inference. The two columns rank the three arms in opposite orders. Pinning the shift at the label drives the residual 8 times lower than the published model above 135 mm and 6 times below it, and leaves the recovered focal *worse*, because a point map with the wrong focal shape plus a compensating shift satisfies it exactly. Solving the shift jointly removes that freedom: the residual is the worst of the three and the telephoto focal is the best. Measured on a fixed random sample of 240 benchmark photographs.

is not: ≥300 mm is the most concentrated and removing its largest contributor *raises* our margin, while 135–300 mm, which looks safer, loses 1.8 points.

F.14 Sensitivity to radial distortion

We assume a centred pinhole camera and ignore radial distortion, which every geometry-based method models explicitly. On MegaDepth with distortion left in, our model does not move and GeoCalib loses 1.7 points (Table 26). This is the expected direction: a method that estimates distortion has something to lose when the assumption it corrects for is violated, and a method that ignores it was never relying on it.

F.15 Seed noise

Three runs per backbone, differing only in seed. The ViT-L we report is reproducible to 0.42 points in aggregate; its per-band figures scatter by up to 2.3, so a per-bucket claim needs roughly five times the aggregate threshold. Table 27 gives all six runs. Every per-band claim in this paper is stated against that spread rather than against the aggregate one.

F.16 Degradation and reframing

Table 28 perturbs the benchmark three ways. Two of them leave the field of view alone and the answer barely moves: re-encoding at JPEG q30 costs 3.7 points and halving the resolution 1.8, neither

run	MRE ↓	δ_{25} ↑	δ_{25} ≥135 mm ↑
AnyCalib, published	36.9	48.2	0.0
18k×8 *(reported)*	28.8	65.4	50.7
18k×8, square training frame	35.2	61.7	51.4
12k×3, square frame	32.6	60.8	43.7
40.6K source, 11k×3	29.7	62.4	40.6
40.6K source, $\sqrt{n}$ weight	33.0	62.1	43.5
40.6K source, telephoto only	191.2	16.1	46.3
40.6K source, panorama crops	37.1	55.4	41.6
18k×8, only ≤110 mm	29.8	58.3	3.4
18k×8, half panorama crops	29.3	64.4	49.1
MoGe-2, published	36.7	47.5	8.2
18k×8, shift joint *(reported)*	30.6	50.7	19.5
12k×2, shift joint	31.1	54.0	28.2
12k×2, shift pinned at the label	58.7	43.8	8.8

Table 24: **Every fine-tuning run, both methods.** The reported rows use one recipe — 18k images, eight epochs, focal bands weighted equally — fixed before either method was run and shared between them. Neither reported row dominates its own block: the AnyCalib recipe leads on mean error and δ_{25} but not δ_{25} above 135 mm; the MoGe-2 recipe leads on mean error but not δ_{25} or δ_{25} above 135 mm. We report every arm anyway: choosing the one that scores highest on the benchmark is the selection this paper objects to elsewhere. For AnyCalib, how hard the sample is reweighted towards the scarce bands is the variable that matters, and the telephoto-only arm is the control showing the cost comes from what was removed rather than from fine-tuning as such. For MoGe-2 the shorter run is better above 135 mm than the longer one while its training objective is worse, which is the residual misalignment of Sec. 6 in a milder form.

band	n	photographers	largest share	margin	without them	Δ
<20 mm	405	169	14%	+9.8	+9.9	+0.1
20-35 mm	636	415	3%	+0.1	-0.2	-0.3
35-70 mm	630	410	6%	-4.3	-4.7	-0.4
70-135 mm	620	380	10%	+5.3	+4.8	-0.5
135-300 mm	483	237	12%	+15.9	+13.7	-2.2
>=300 mm	403	173	25%	+17.4	+18.4	+1.0

Table 25: **The margin over the best published method when each band loses its largest contributor.** Photographs by one person share a camera, a lens set and an EXIF pipeline, so they are not independent draws and every per-image interval in this paper is correspondingly too narrow. Concentration and dependence turn out to be different things: the ≥300 mm band is the most concentrated, one contributor holding a quarter of it, and removing them *raises* our margin. The band that actually rests on one photographer is 135-300 mm, where the margin falls by 2.2 points. One band changes sign: 20-35 mm from $+0.1$ to -0.2, within two s.d. of that band's seed variation (s.d. 0.9) — a tie either way.

biasing the prediction. The third is the informative one. A centre crop to a fraction c of the frame multiplies the true 35mm equivalent by $1/c$, so a correct model must change its answer, and ours tracks 46 to 59% of that factor. The test does not separate reading perspective from recognising scenes: MoGe-2 tracks about 100% of the factor and UniDepthV2 about 88%, yet both collapse above 135 mm. What the published methods lack is absolute scale at long focal lengths, not sensitivity to relative change. The stronger evidence for our model is optical zoom with fixed content, where it follows InFlux's zoom with a slope of 0.99 (Appendix F.1).

F.17 Scaling by focal band

Across a ladder from 781 to 89,651 training images, three seeds each on a ViT-S, the band-balanced mean error falls from 44.7% to 27.9%, close to a power law in the number of images with exponent -0.097. The scarce ≥300 mm end falls 56.9% to 30.6% in two regimes, steeply below 12,500

method	pinhole	radial	Δ
Depth Pro	13.1	12.9	-0.2
AnyCalib	15.1	15.3	+0.2
MoGe-2 + our data†	17.1	17.4	+0.3
Ours (ViT-L/14)†	12.0	12.4	+0.4
AnyCalib + our data†	14.9	15.5	+0.6
UniDepthV2	10.1	10.8	+0.7
WildCamera	14.2	15.6	+1.4
MoGe-2	13.7	15.3	+1.6
GeoCalib	20.9	22.6	+1.7

Table 26: **Sensitivity to radial distortion**, on MegaDepth with and without it (mean relative error, %), sorted by how much each method moves. We assume a centred pinhole camera, which reads as a limitation and measures as a mild advantage: the methods that estimate distortion have something for it to break, and a regressor over semantic features has less. Every row moves by under a point except WildCamera, MoGe-2 and GeoCalib, so the effect is real but small, and we report it as a property of the formulation rather than a result we tuned for. Rows are ordered by how far they move and none is marked best: each column is one run on a split with no measured seed spread, so a 0.2-point separation is not one we can resolve.

seed	MRE ↓	MedRE ↓	δ_{25} ↑	<20	20-35	35-70	70-135	135-300	>=300
ViT-L, 1	19.60	10.44	74.9	20.0	17.4	18.7	18.4	21.3	24.0
ViT-L, 2	19.19	10.78	75.1	17.9	16.2	18.4	18.3	21.0	25.6
ViT-L, 3	19.29	10.86	75.0	18.1	15.6	18.6	19.1	21.5	25.0
s.d. ViT-L	0.21	0.22	0.10	1.13	0.90	0.15	0.44	0.27	0.80
ViT-B, 1	22.35	13.02	70.1	22.0	17.2	21.7	24.1	23.1	28.1
ViT-B, 2	22.31	12.27	71.0	21.7	20.0	21.7	21.3	24.4	26.6
ViT-B, 3	22.37	12.28	71.3	20.8	20.2	22.3	21.3	23.4	27.8
s.d. ViT-B	0.03	0.43	0.63	0.65	1.64	0.36	1.62	0.66	0.76

Table 27: **Seed noise, measured where it is used.** Three runs per backbone on the full training split, differing only in seed. The ViT-L we report varies by 0.21 points (one s.d.) in aggregate, so aggregate differences of half a point are outside it — but its per-band figures scatter by up to 1.1, so a per-bucket claim needs roughly five times the aggregate threshold.

images and half as fast above, while 135–300 mm gets worse, from 24.6% to 30.0%: an aggregate over mixed focal content hides effects of opposite sign, which is also how the collapse escaped the benchmarks of Sec. 3. Table 29 is the numeric form of Fig. 10, with all six bands rather than the three the figure plots. The ≥300 mm band falls monotonically across the whole sweep, by 26 points and with no plateau, and the wide end falls by 34, while 135–300 mm is the one band that gets *worse*, by 5.4. The bands do not share a shape, so the aggregate curve, which is what a single mean would show, describes none of them and understates the ends.

F.18 Dataset size at fixed architecture

Two dataset sizes on two backbones, with architecture, schedule and seed held fixed. It is the two-point version of Fig. 10, kept because it is the only place the ViT-L appears on a scaling comparison; the eight-point curve is measured on ViT-S.

F.19 Architecture ablations

The reported models train for 25 epochs at batch 64 with AdamW, at 10^{-3} for the head and 10^{-5} for the backbone. Unfreezing depth is the dominant architectural factor and gives an inverted U on both backbones. The *slope* column is the gradient of a linear fit of $\log \hat{f}$ on $\log f$; below one means the model hedges towards the centre of the range, and it rises towards one as accuracy improves. Each row is a single run.

perturbation	MRE ↓	Δ	pred. ratio	expected	tracked
unperturbed	19.60	+0.00	1.000	1.000	—
JPEG q75	20.49	+0.89	0.998	1.000	—
JPEG q50	21.93	+2.34	1.001	1.000	—
JPEG q30	23.29	+3.70	1.005	1.000	—
rescale 0.5x	21.40	+1.80	1.025	1.000	—
rescale 0.7x	20.71	+1.11	1.018	1.000	—
rescale 1.5x	19.84	+0.24	1.000	1.000	—
centre crop 0.9	21.27	+1.67	1.052	1.111	46%
centre crop 0.8	23.70	+4.10	1.136	1.250	54%
centre crop 0.7	25.83	+6.23	1.251	1.429	59%

Table 28: **Degradation and reframing.** JPEG and rescaling do not change the field of view, so the answer should not move, and it largely does not: the worst case is heavy JPEG at q30, costing 3.7 points without biasing the prediction. A centre crop to a fraction c of the frame multiplies the true 35mm equivalent by $1/c$, so here the answer *must* move; the last column is how much of that the model actually does. It tracks roughly half, which is evidence it reads perspective rather than recognising the scene — a model doing the latter would track nothing — and equally evidence that it under-corrects, which is why the error still rises. Errors are against the crop-adjusted target. Reported ViT-L/14, 3,177 images.

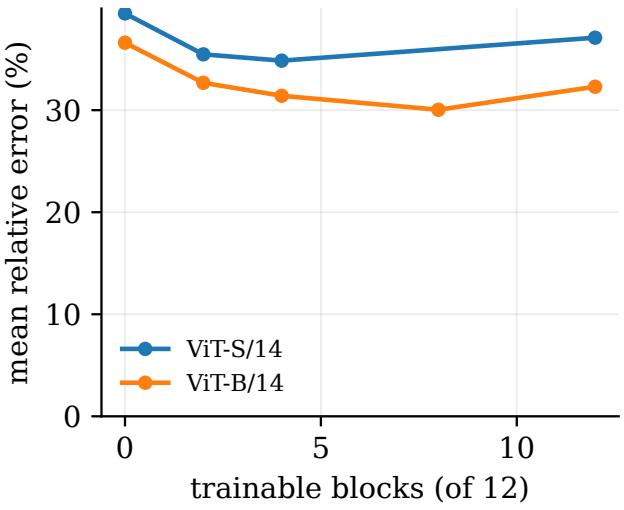


Figure 9: **Unfreezing depth gives an inverted U** on both backbones, so the common frozen-versus-full comparison samples only the two endpoints. The sweep covers ViT-S and ViT-B; the ViT-L depth we report, 16 of 24, is chosen by the same fraction rather than swept, so it is an extrapolation from these two curves.

F.20 What the training distribution looks like

The released set is flatter in $\log f$ than its predecessors. Effective focals is the perplexity of the distribution over half-stop bins, the same statistic Table 1 reports: 7.8 at 38.8K, 9.4 at 50.9K, 10.9 at 100K, while the largest single bin falls from 28.2% to 19.4%. The set did not merely grow; its mass moved out of the normal-lens peak and into the tails.

F.21 Input resolution

Every number we report comes from a model that sees 252 px, while the published methods resize the same input to their own native resolutions, several of them larger. FocalBench ships its images at most 512 px on the long edge (590 of them, copies from the MegaScenes mirror, at 320 px), so that is the detail available to every method including ours: a method whose native size exceeds it

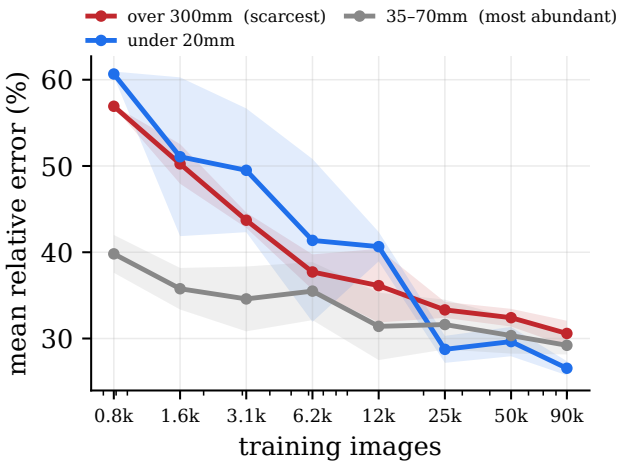


Figure 10: **What more data buys, and where.** Eight sizes on a halving ladder from 781 to 89,651 images, three seeds each, architecture and schedule fixed. The bands do not share a curve: the scarce ≥300 mm end falls steeply and is still falling at the full set, the abundant 35–70 mm middle barely moves, and 135–300 mm gets *worse* , the range trade of Sec. 6 seen from the data side. ViT-S/14, 4 of 12 blocks unfrozen: evidence about the data, not about the model we report.

bucket	0.8k	1.6k	3.1k	6.2k	12k	25k	50k	90k	Δ	s.d.
<20	60.7	51.1	49.5	41.4	40.7	28.7	29.6	**26.6**	-34.1	1.7
20-35	48.9	39.6	35.1	34.0	30.7	27.4	26.3	**23.5**	-25.4	1.4
35-70	39.8	35.8	34.6	35.5	31.4	31.6	30.3	**29.2**	-10.6	2.6
70-135	37.6	37.2	35.6	35.4	32.6	30.7	29.1	**27.3**	-10.3	1.8
135-300	**24.6**	29.6	32.7	36.0	34.8	33.8	31.6	30.0	+5.4	0.8
>=300	56.9	50.2	43.7	37.7	36.1	33.3	32.4	**30.6**	-26.3	1.2
balanced	44.7	40.6	38.5	36.7	34.4	30.9	29.9	27.9	-16.9	–

Table 29: **Scaling by focal bucket.** Three seeds per point; s.d. is the typical spread across them for that band, so a Δ within twice it is not a change. The balanced row is the mean over bands, which is a power law in $\log n$: slope -0.097, $R^2 = 0.99$ over $115\times$. The bold entry is the best training-set size for that band. The ≥300mm bucket falls monotonically by 26 points with no plateau, while 135-300 is the one band that gets *worse*, by 5.4. An aggregate mean over these mixes the two and understates the effect the dataset was built to produce.

is upsampling, not seeing more. That makes our comparison conservative rather than unfair, but it leaves open whether 252 px is what bounds us at the long end, where discriminating 600 mm from 900 mm is a question about fine detail. Table 32 answers it: at matched epochs on the same images, 504 px is *worse*, by 1.1 points of mean relative error overall and 7.7 at or above 300 mm, on a ViT-B trained on the earlier 38.9K set rather than the reported recipe. Only the wide end improves; the two bands beside it are ties at this spread, and every band above 70 mm is worse. In this single comparison, higher resolution did not help at the long end.

backbone	images	MRE ↓	MedRE ↓	δ_{25} ↑
ViT-B/14, 8 of 12	29,761	28.3	17.4	61.8
	40,623	24.5	15.0	65.9
	89,651	**22.4**	**13.0**	**70.1**
	Δ	-5.9	-4.3	+8.3
ViT-L/14, 16 of 24	29,761	25.7	14.4	65.8
	40,623	21.1	11.9	70.8
	89,651	**19.6**	**10.4**	**74.9**
	Δ	-6.1	-4.0	+9.1

Table 30: **What the data is worth.** Identical architecture, schedule and seed; the only variable is how many of our images the model saw. Both backbones gain at least 5.9 points, far above either backbone's own seed variation (s.d. 0.21 for the ViT-L, 0.03 for the ViT-B; Table 27), and a ViT-B on the full set matches a ViT-L on the 29.8K set. The versions are not random subsets of one another: each is larger and flatter in $\log f$ than the last. Counts are the training split; the release also holds out 6,700 images for the benchmark pool and 3,630 for validation. Δ is the last row minus the first, so a gain is negative on the two error columns and positive on δ_{25}.

backbone	trainable blocks	MRE ↓	MedRE ↓	δ_{25} ↑	slope
ViT-S/14	2 of 12	35.5	24.4	50.8	0.718
ViT-S/14	4 of 12	34.8	22.0	54.1	0.740
ViT-S/14	all 12	37.1	22.9	52.7	0.725
ViT-S/14	2 of 12, binned head	40.4	25.4	49.3	0.713
ViT-B/14	2 of 12	32.7	21.0	56.2	0.749
ViT-B/14	4 of 12	31.4	20.6	57.0	0.756
ViT-B/14	8 of 12	30.0	18.7	58.5	0.763
ViT-B/14	all 12	32.3	19.2	58.7	0.759
ViT-L/14	16 of 24	**26.2**	**15.5**	**64.2**	0.802
ViT-S/14	frozen	39.5	28.3	45.2	0.649
ViT-B/14	frozen	36.6	25.8	48.6	0.698
ViT-S/14	frozen, binned head	40.1	28.9	44.6	0.659
ViT-S/14	frozen, binned, lr 3e-4	41.4	28.1	45.5	0.680
ViT-B/14	frozen, binned head	37.9	26.9	47.2	0.691

Table 31: **Unfreezing depth is the dominant architectural factor.** Both backbones give an inverted U: error falls as blocks are unfrozen and rises again when everything is, and the rise at the end is 2.3 points on both — far outside the seed band. Where exactly each peaks is a weaker claim and we do not make it: the ViT-S/14 step from 2 to 4 blocks is 0.7 points from single runs, and no seed spread is measured for ViT-S at this size, so the shape is resolved and the position of its minimum is not. The four binned-head arms are each worse than the MLP run with the same backbone and unfreezing depth, which is the evidence for that claim in Sec. 6. The log-space slope rises with accuracy, so the model hedges less as it becomes better informed; it is a diagnostic rather than a score, so it is not bolded. Every row is a single run.

input	MRE ↓	δ_{25} ↑	<20	20-35	35-70	70-135	135-300	>=300
252 px	**29.7**	**59.5**	31.6	**26.4**	**24.6**	**24.2**	**32.1**	**46.3**
504 px	30.8	56.8	**29.7**	**25.9**	**24.2**	**25.3**	34.4	54.0

Table 32: **Input resolution, at matched epochs.** ViT-B/14, 8 of 12 blocks unfrozen, 25 epochs both, on the same 38,855 images in the same order, so resolution is the only variable. Doubling it costs 1.1 points of MRE overall and 7.7 at or above 300 mm: more pixels do not buy telephoto accuracy, and at equal epochs they cost it. The wide end is where the extra detail helps, which fits perspective distortion being the cue there. This is the 38.8K subset, the only one with a 512 px cache, so it bounds the resolution question rather than restating the headline model. Bold marks the better value and anything tied with it, within two s.d. of the ViT-B seed variation for that column, so 20-35, 35-70 and 70-135 read as the ties they are: two single runs cannot separate 1.1 points.

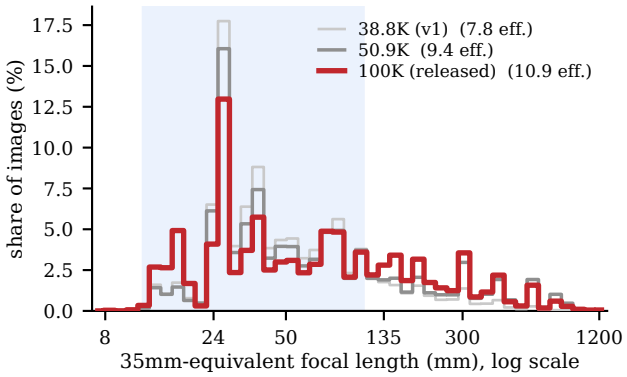


Figure 11: **The released training set in** $\log f$, against the two smaller versions it replaces. Shaded is the range the panorama generator behind GeoCalib and AnyCalib can produce. Each successive version flattens the 24–35 mm peak and adds mass beyond 135 mm, which is what the per-band scaling in Table 29 responds to.