

A Calibration Audit of Confidence in Feed-Forward 3D Reconstruction

Nanxing Nick Deng^{1,*}

Qing Cheng^{1,2,*}

Niclas Zeller³

Daniel Cremers^{1,2}

¹Technical University of Munich

²Munich Center for Machine Learning (MCML)

³Karlsruhe University of Applied Sciences

*Equal contribution.

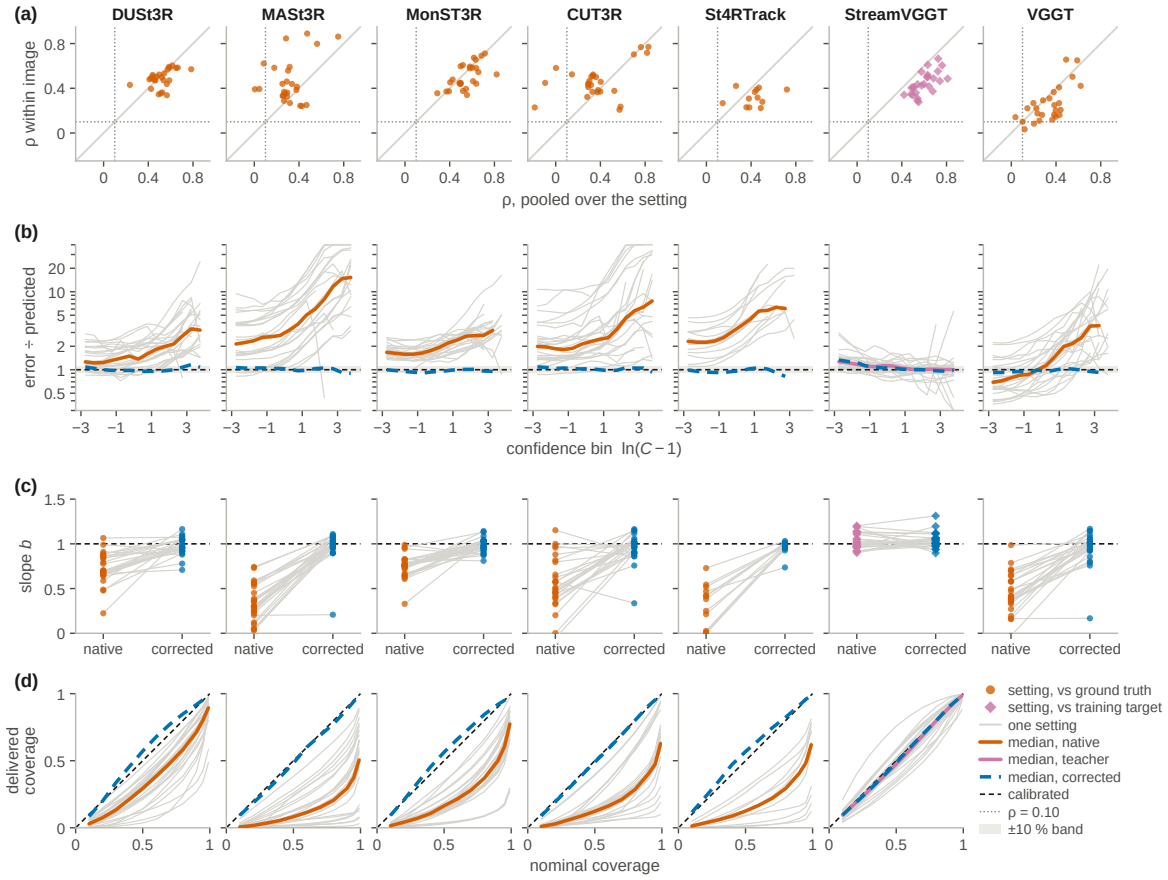


Figure 1. An audit of native confidence over seven backbones and 13 datasets, and the effect of a post-hoc correction. A *setting* is one backbone’s output head on one dataset. The grid is 168 settings, each scored at full pixel population. The correction is a two-constant power law in the predicted uncertainty (Sec. 6), fitted on a calibration half of each setting’s scenes. In (d) it is scored on the other half, in (b) and (c) it is applied to the setting’s full profile. (a) Ranking of predicted uncertainty against error. (b) Level, mean error over predicted uncertainty per confidence bin. (c) Slope of the calibration curve, which equals 1 when predicted uncertainty grows at the same rate as error. (d) Delivered against nominal coverage, at eleven nominal levels. StreamVGGT is read against its training teacher. VGGT’s hyperparameter α is unattested, so its coverage is not drawn.

Abstract

Feed-forward 3D reconstruction models emit a per-pixel confidence that downstream systems read as a reliability

signal. It is trained as a loss weight, not as an uncertainty magnitude, and whether it can be used as an error prediction has not been measured. We audit seven released backbones on thirteen datasets and score the confidence on four

properties, how well it ranks error, whether its level is right on average, whether it holds across the confidence range, and whether its intervals cover the truth. The confidence ranks error well, but the predicted uncertainty is too low when it is read under conditions that are not exactly those of training. The median case is off by $2.4\times$ across all seven models, and the error prediction is further off the more confident the model is. We show that this phenomenon can appear even though the loss’s optimum is reached. A released model resumed under its own loss reaches that optimum on its training data within a few hundred updates and stays overconfident on unseen frames. A power law with two constants per backbone and dataset corrects the overall magnitude of the predicted uncertainty and leaves the ranking untouched. What no rescaling reaches is the scene, which we attribute to the model’s missing knowledge of scale across predictions. Every correction we tried is close to right on average and still leaves two thirds of held-out scenes outside a five-point band, because what a scene is missing is a shape rather than a shift. We release the audit protocol, its results, and the fitted constants per model and dataset. Fitted with the target dataset held out, the constants bring the median case from $2.4\times$ off to $1.35\times$, and a refit on a few labelled scenes of that dataset reaches $1.12\times$.

1. Introduction

Feed-forward models reconstruct 3D geometry from uncalibrated images in a single pass and emit a per-pixel confidence alongside it [36, 55–57], a lineage surveyed in [61]. That confidence is the only reliability signal these models ship, and downstream systems use it. They threshold points before fusion, weight residuals in bundle adjustment, or set the measurement noise of a filter.

It is trained as a weight in a confidence-weighted regression loss, not as an error bar, and practice already reflects the difference. SLAM3R thresholds two DUST3R-style heads at 1.5 and at 12 [39]. FILT3R runs a Kalman filter on a frozen CUT3R, adapts the process noise, and leaves the measurement noise a fixed scalar [23]. DeepFactors computes a weight from a predicted standard deviation and then returns one instead [10]. The one visual-inertial system we found that does use a learned uncertainty as a covariance first fits a constant gain on a validation set, so that the mean Mahalanobis distance comes out at one [25]. Every one of these choices avoids relying on the confidence’s level. It is consumed as a ranking, and where a level is needed it comes from elsewhere. Whether that caution is justified, and by how much, has not been measured.

The field’s answer has been to replace the head. Evidential regression [1, 62], heteroscedastic regression [29, 41], MC dropout [13] and deep ensembles [32] all produce a predictive distribution, and recent work reports large gains

in confidence ranking on these backbones [62]. That is the right answer when retraining is possible. Two things remain. Released checkpoints still ship the original head, and those checkpoints are what systems are built on. And the problem is not an undertrained head. We resume a released model under its own loss, and it reaches that loss’s optimum on the data it is training on within a few hundred updates. A head retrained the same way is overconfident on frames it has not seen.

We therefore ask what the native confidence is worth as a level, and how much of the gap a user can close without touching the weights. The loss is minimised where the confidence equals a constant over the expected residual (Sec. 3), so inverting the trained relation returns the mean residual of training, measured against the target the loss used, over the data it was trained on, in the units it was written in. A deployed model is read against ground truth, on new data, in the scene’s own units. We measure the resulting error on 168 settings (Fig. 1) over seven released backbones and thirteen datasets at full pixel population (2.786×10^{10} scored pixels). We then measure how much of it post-hoc recalibration recovers.

Contributions.

- An evaluation protocol and the grid it runs on, the paper’s main resource. A released confidence is scored on four properties at once, ranking, level, slope, and coverage, whereas this literature scores ranking alone (Sec. 4, Tab. 1).
- What it reads (Fig. 1, Tab. 2). The confidence ranks error well and understates the error by a median factor of 2.4 across the grid, and by more where it is most confident. Sec. 5 shows this is where the loss’s own optimum lies, not where training happened to stop.
- A study of the corrections a user of a frozen model can actually apply. Two constants per backbone and dataset remove most of the level error and leave the ranking untouched. The finer per-bin correction the literature would reach for costs ranking and buys little. None of them reaches the scene. How far one scene sits from another is set only by that scene’s own labels (Sec. 6).
- That correction pre-fitted for every backbone and dataset, so a reader with no labelled data can apply it to a frozen checkpoint (Tab. 4).

Our scope is post-hoc recalibration of the frozen native confidence. We do not claim that a head trained from scratch cannot rank better. Published work reports that it does [62]. We do not measure the effect of either reading on a downstream system.

2. Related Work

Dense uncertainty and how it is scored. Per-pixel uncertainty in dense prediction is conventionally scored by

how well it ranks error. Sparsification error and its summary AUSE come from optical flow [22] and are standard in monocular depth [46]. Stereo confidence is scored the same way, through ROC curves [47]. Whether the number is the right *level* is rarely asked. The exceptions change the training objective to make it so [8, 40], while label-free post-hoc methods improve ranking instead of correcting a scale [16, 21]. One independent study reaches our direction from outside, reporting that VGGT’s depth confidence “exhibits the typical phenomenon of overconfidence” on DTU [20], though on a scale-free proxy, so it corroborates the sign and not the size.

Replacing the head. Recent work replaces the loss weight with a head that is a predictive distribution by construction. The options are evidential regression with a closed-form Student- t posterior [1, 62], heteroscedastic regression [29, 41], MC dropout [13] and deep ensembles [32]. Trust3R reports 25 % lower AURC and 41 % lower AUSE on ScanNet++ [62]. We take that as settled and disjoint from what we measure, since those are ranking metrics on a retrained head and ours are level metrics on the native one. Whether a *differently* trained head has the right level is an open question we can only bound from outside the family. Sec. 5 scores an uncertainty read off sample disagreement and a direct error regressor on the same grid, and both miss, in opposite directions.

Post-hoc recalibration. Rescaling a frozen model’s uncertainty is standard [18, 31, 37, 45]. The published form for regression is a single global scalar fitted by likelihood on a labelled validation set [33]. The same two corrections appear in the conformal literature as split conformal prediction [35, 43] and its per-group refinement [5, 54]. What such a guarantee can be conditioned on is settled theory [4, 34, 54], and the methods that pursue conditional coverage where it is attainable act on pre-specified groups [24], kernel neighbourhoods [17], a covariate shift between calibration and test [51], or a target stated as a risk [3]. Sec. 6 runs the first four, and the risk-control target is in the supplement. Conformal constructions have reached dense per-pixel regression [2] and geometric estimation [58], but not, as far as we know, monocular depth or feed-forward pointmap reconstruction.

Uncertainty in geometric systems. The loss we analyse descends from learned loss attenuation, where a predicted variance down-weights hard pixels during training [29]. That a frozen geometric system’s uncertainty is systematically wrong and correctable offline has precedent in visual-inertial odometry [52], and overconfidence has been diagnosed in semantic 3D mapping and traced to the fusion rule [9].

3. Confidence as a Loss Weight

Loss. The DUST3R family [57] predicts a 3D point and a confidence C_i at every pixel i , trained with a confidence-weighted regression loss [12, 36, 56, 60]

$$L = \sum_i (C_i e_i - \alpha \log C_i), \quad (1)$$

where $\alpha > 0$ is a fixed hyperparameter and e_i is the regression residual, the distance between the predicted 3D point and the point the loss scores it against. The first term weights the residual by its confidence, and the second stops the confidence collapsing to its floor. Since the head emits $C = 1 + e^x$, $C > 1$ and the decoded $\hat{\sigma}$ is bounded above by α (supplement, §4). Residuals are normalized during training by a scale computed on each forward pass, so they are fractions of the size of the reconstruction and not lengths.

Optimum. Setting $\partial L / \partial C_i = 0$ gives $C_i = \alpha / e_i$. The head maps a per-pixel feature to a point and a confidence, so pixels whose features look alike receive the same confidence while their residuals differ. Since L is linear in e_i , the expected loss over pixels sharing a confidence depends on their residuals only through the mean and is minimized at $C_i = \alpha / \mathbb{E}[e_i]$. Reading the confidence as an uncertainty means inverting the trained relation,

$$\hat{\sigma}_i = \alpha / C_i, \quad (2)$$

which at the optimum returns $\mathbb{E}[e_i]$, the mean residual over training pixels of the same confidence. We hold α at the value each release attests, 0.2 for the DUST3R family and 0.1 for StreamVGGT, and never fit it ourselves. VGGT is the one backbone whose α is only a function default, and Sec. 5 reads it accordingly. With $z = e / \hat{\sigma}$, the residual in units of the predicted uncertainty, the optimum becomes the statement

$$\mathbb{E}[z \mid C] = 1, \quad (3)$$

which follows from Eq. 1 alone. Read as a likelihood the loss also implies a distribution for e , as a learned variance does in heteroscedastic regression [29], but its optimum fixes only the mean.

Three conditions. The quantity Eq. 2 returns is a conditional expectation, and training fixes the three quantities it is taken under. The residual is measured against *the loss’s target*, ground truth for most audited models and a teacher’s prediction for one. The expectation is over *the training distribution* as the model’s own pipeline samples it. And the residual carries *the loss’s normalization*. A reader evaluates against ground truth, on data of their own, in the scene’s own units, so all three can differ. Sec. 4 matches the first and third to training wherever the release permits.

Property	What it decides for a user	How we score it
Ranking	Which pixels to keep or drop, and any percentile cutoff. Invariant to rescaling.	Spearman ρ between $\hat{\sigma}$ and the error, pooled and within images.
Level	Whether an interval covers, and whether a cutoff in absolute units means what it says.	$\bar{u}_j$, delivered error over predicted, per confidence bin (Eq. 4).
Slope	Whether <i>relative</i> weights are right. A least-squares weight $\hat{\sigma}^{-2}$ over-weights confident pixels when $b < 1$.	b in Eq. 5, the calibration curve fitted across a setting’s bins.
Coverage	Whether a stated interval holds, on average and on the scene at hand.	Fraction of truth inside $\hat{\sigma} m_p$, at eleven nominal levels, per scene and pooled.

Table 1. The four properties we score, and what each one decides. They are distinct, and Sec. 6 reports all four for every correction.

4. Audit Protocol

Setup. We decode the uncertainty as $\hat{\sigma} = \alpha/C$ and match the target and the normalization to the ones each model trained with, wherever the release permits (supplement, Tab. S1). A *setting* is one backbone’s output head on one dataset, and each is scored on the four properties of Tab. 1. They answer different questions, so we never summarize them into one number.

Grid. Six of the seven backbones expose two heads and one a single head, giving $26 \times 6 + 12 = 168$ settings over 13 datasets. These are ScanNet (val, train) [11], TUM (static, dynamic) [50], 7-Scenes [49], BlendedMVS [59], Bonn [42], KITTI [14, 53], CO3Dv2 [48], DL3DV [38], Sintel [6], Virtual KITTI 2 [7] and Dynamic Replica [26], the same for every backbone. We sample eight-frame clips, so scene, clip and frame are three nested units we can read the same statistic on. Some datasets are in a model’s training diet, but the pixels scored are never those training saw. Three of the 13 are close relatives of others, so they amount to ten independent sources, and one backbone cannot run CO3Dv2. Every setting is scored at full pixel population, 2.786×10^{10} pixels, a pixel counting when its ground truth is valid, its depth is in range, and it falls inside the model’s field of view.

Metrics. Ranking is scored by the Spearman correlation ρ between $\hat{\sigma}$ and e , with a setting passing at $\rho \geq 0.10$. We report two, on different populations. One pools every scored pixel. The other correlates *inside each image* and takes the median over the setting’s images, which is what rules out a confidence tracking only the per-image error scale. The second runs on fewer pixels, dropping the confidence floor and the out-of-range bins, a median of 7 % of a setting and 81 % on the worst, so the two are not one statistic at two granularities.

Confidence bins. Pixels are assigned to 14 bins fixed on intervals of $x = \ln(C - 1)$, since $C = 1 + e^x$. With $\mathcal{B}_j$ the j -th bin and n_j its pixel count, the calibration level

$$\bar{u}_j = \frac{1}{n_j} \sum_{i \in \mathcal{B}_j} z_i \quad (4)$$

is the factor by which delivered error exceeds predicted uncertainty in that bin. Eq. 3 requires $\bar{u}_j = 1$ everywhere. Intervals on $\bar{u}_j$ come from resampling *scenes*, not pixels. Pixels of one scene share a camera, a depth range and a normalization, so they are not separate evidence. A bin fails high when its interval lies entirely above 1.10 (over-confident) and low when entirely below 0.90, counted over bins holding at least three scenes and outside KITTI’s pose-contaminated block. Write $\bar{e}_j = \bar{u}_j \hat{\sigma}_j$ for the mean delivered error, and $\hat{\sigma}_j$ for the predicted uncertainty at the bin midpoint. We fit the calibration curve

$$\ln \bar{e}_j = c + b \ln \hat{\sigma}_j \quad (5)$$

across a setting’s bins, weighted by pixel count. The intercept absorbs any overall factor, the level again, so b measures how the bins differ. Correct uncertainties give $c = 0$, $b = 1$. A slope $b < 1$ means the predictions are spread wider than the errors.

Intervals. The level fixes the mean of z in a bin and leaves the rest of its distribution open. Coverage asks whether $\hat{\sigma}$ can serve as that distribution’s scale. If it can, the truth lies within $T_p = \hat{\sigma} \cdot m_p$ of the prediction a fraction p of the time, m_p being the level- p quantile of a mean-one model for z . We take m_p either as the quantile of a mean-one Gamma fitted to the within-bin residual, or as the empirical quantile of z inside the pixel’s own bin, which assumes no distribution at all. Each correction is scored under whichever its construction implies. Both are fitted on a calibration half of a setting’s scenes and counted on the held-out half, one fixed split per setting, at eleven nominal levels from 10 % to 99 %. We read two numbers off this. *Delivered* coverage is the share of truth inside the interval, pooled over the pixels of one scene. *In band* is the fraction of a setting’s held-out scenes whose own delivered coverage falls within 0.05 of nominal. Pixels are pooled inside a scene and never across scenes, as the bootstrap above resamples scenes, and a grid figure is a median over settings of a per-setting number. We do not fold these into a single calibration error [31, 37] or report a likelihood, since the loss leaves the residual distribution open. The supplement gives the per-backbone channel, target, normalization and α , compares ρ against the sparsification family [22, 30], and fits the residual shape.

5. Calibration Audit

Ranking. The confidence ranks the error almost everywhere. Within images, ρ is positive on all 168 settings and

Backbone	#	Ranking ρ		Failing bins		Level, slope, coverage		
		pooled	image	high	low	low $\rightarrow$ top bin	b	cov
DUST3R	26	0.52	0.49	208	0	1.26 $\rightarrow$ 3.22	0.70	0.71
MonST3R	26	0.55	0.47	308	0	1.67 $\rightarrow$ 3.19	0.76	0.50
MASt3R	26	0.29	0.39	293	0	2.14 $\rightarrow$ 15.24	0.30	0.29
CUT3R	26	0.33	0.45	290	0	1.99 $\rightarrow$ 7.62	0.50	0.40
St4RTrack	12	0.44	0.31	125	0	2.31 $\rightarrow$ 6.08	0.40	0.40
All five	116			1,224	0			
StreamVGGT	26	0.59	0.43	40	10	1.26 $\rightarrow$ 1.01	1.01	0.90
vs. GT				297	0	2.47 $\rightarrow$ 3.35		
VGGT	26	0.37	0.21	120	64	0.69 $\rightarrow$ 3.67	0.42	
Grid	168	0.46	0.43				0.65	

Table 2. The audit by backbone, medians over its settings. # is the number of settings, image the within-image ρ , low $\rightarrow$ top bin the level $\bar{u}$ of Eq. 4 at the lowest and at the highest bin supported by at least half of the backbone’s settings, and cov the delivered coverage at nominal 0.90. The five rows above the rule share a referent and sum to the subtotal. A bin fails high when its confidence interval lies entirely above 1.10 and low when entirely below 0.90. KITTI’s pose-contaminated bins are excluded, 20 across the five. The counts are a census over 90 backbone-dataset pairs, not independent tests. StreamVGGT is read against its teacher and the indented row against ground truth. VGGT’s α is a function default, so its bins are reported and not counted.

above threshold on 166 (Fig. 1a, Tab. 2). The ranking runs backwards on about one image in twelve at the median setting. Seven settings fail when pixels are pooled across images while passing within them, all on CO3Dv2 or BlendedMVS.

Level. The level is another matter. On the five backbones trained against ground truth, every confidence bin that fails does so by claiming too little error, 1,224 of them, and not one the other way. Across the grid the median setting is off by 2.4 $\times$, across these five by 2.8 $\times$. StreamVGGT, trained against a teacher and read against that teacher, is the one backbone close to calibrated, though on that target it still fails high on 40 bins and low on 10 (Tab. 2). Read against ground truth, it joins the others. VGGT’s α is a function default, so its bins are reported and not counted. The five that remain descend from one loss, so the count is a census over a single lineage and not five independent tests. Fig. 2 shows one held-out frame of two settings, where the uncertainty tracks the error’s structure and sits below it almost everywhere.

Slope. The predictions are also spread wider than the errors (Fig. 1c). The calibration curve’s slope b lies entirely below one on 123 of 168 settings. The predicted error at a setting’s 90th percentile of confidence is about a sixth of that at its 10th percentile. A slope below one means the model grows *more* overconfident as its confidence rises, so the predictions a reader would weight most heavily are the

ones furthest off. Part of the flatness is aggregation rather than a flat head, since a setting whose scenes are each calibrated at their own level still fits at 0.93. Two checks bound that. Restricting to the ground-truth settings leaves 114 of 142 below one, and fitting inside single frames, where no aggregation is possible, puts the slope below one on all 54 settings of the four backbones where the fit is available. The headline count is the grid one, 123 of 168, and the other two say the effect survives its most likely artifact.

Coverage. On the five backbones trained against ground truth the same overconfidence appears in the intervals, which are too narrow at all eleven nominal levels and proportionally worst at the tight end, where MASt3R’s nominal 10 % interval holds 0.6 % of the truth (Fig. 1d).

Domains and units. Sorting datasets by indoor or outdoor, static or dynamic, real or synthetic ground truth moves the level by at most 1.6 $\times$, and datasets outside every training diet read no worse than those inside. What moves it is the normalization a dataset was trained under. Read in metres, CUT3R is 3.0 $\times$ off on the datasets it trained on with metric residuals and 15.6 $\times$ on those with scene-normalized ones. The spread across a backbone’s datasets is that, not scene content, and the supplement gives the level of every setting and marks its branch.

Convergence. A head at the optimum of Sec. 3 would satisfy Eq. 3, so every confidence bin would deliver the error it predicts. No backbone does. Either training never reached that optimum or one of its three conditions moved, and the first is directly testable (supplement). A fresh confidence head on DUST3R’s trunk, retrained on two of DUST3R’s own datasets, is calibrated on the frames it trains on and overconfident on frames it does not see. The released DUST3R put back into training under the same loss behaves the same way. Read on CO3Dv2, a dataset in its diet, it starts 2.79 $\times$ off. That reading is already under changed conditions, since the clips, crops and sampler are ours and not the release’s, and the pixels are ones training never saw (Sec. 4). It reaches 1.10 $\times$ after roughly 300 updates on that draw, and stays within 0.91 to 1.12 $\times$ over ten further epochs under both learning-rate schedules. So the loss reaches its optimum on whatever draw it is currently training on, and the level a reader sees is set by which draw that is rather than by how long the model trained. A backbone trained further under the same loss would, on this evidence, stop at the same place. The miss is in the loss’s target, normalization and training population, not in how far the checkpoint sits from its own optimum. On this evidence a few hundred updates under the same loss on labelled scenes of the target domain would repair the level there, though we ran it only on training data. Sec. 6 is what a reader can do without it.

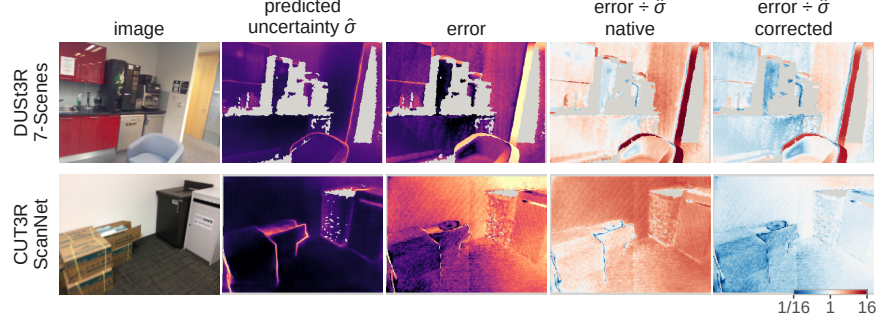


Figure 2. One held-out frame of two settings, chosen at each setting’s median ratio of error to predicted uncertainty $\hat{\sigma}$. The uncertainty follows the error’s structure and sits below it almost everywhere. Right, that ratio before and after the released correction, one pair of constants per setting and not per frame, on a log scale centred at one, red above. Grey is unscored.

The three conditions in isolation. Each of the three conditions moves the reading on its own, with the weights untouched (supplement). Scored against ground truth instead of the teacher it trained on, StreamVGGT goes from calibrated to $2.2\times$. Over four high-confidence bins of MAST3R’s anchor dataset the delivered error is 2.5 to $5.9\times$ the prediction in metres but only 1.1 to $2.1\times$ as a fraction of the scene, so the normalization carries roughly 55 to 64 percent of the overconfidence *there*, one backbone at one anchor and not a family figure. MonST3R and St4RTrack read 1.38 and 0.98 on the scenes they trained on and 2.34 and 6.05 on unseen scenes of the same datasets, so what changes is which scenes, not which domain. DUST3R reads no less overconfident where ground truth is exact, and a pre-registered noise budget can account for at most 12.7% of the failing bins, 158 of the 1,244 counted before the KITTI exclusion of Tab. 2 (supplement), so ground-truth quality is not the cause.

Other designs. The failure is not specific to Eq. 1. Marigold [27, 28] has no trained confidence head and reads its uncertainty off the disagreement among ten sampled predictions, and that uncertainty is too small on all 13 datasets. At nominal 90% it delivers 17 to 41%, against the 79% that ten samples can honestly reach. UniDepth v2 [44] regresses the error level directly and fails the opposite way, predicting three to thirty-four times more error than it delivers on the ten datasets where its metric readout is valid (supplement, Fig. S7).

6. Post-hoc Recalibration

The correction class. The weights are frozen, so the only thing left to adjust is the reported uncertainty. A correction rescales it as a function of the confidence value and the dataset. Every published rescaling of a frozen model’s uncertainty is in that class. The conformal methods that condition on more re-choose the interval instead (Tab. 3,

middle block). One scalar for the whole dataset is split conformal prediction [35, 43], and it is what temperature scaling [18] and σ -scaling [33] do here. Splitting the pixels by confidence bin and giving each bin its own scalar is the same correction at a finer grain [5, 54]. Nothing else about a pixel enters a rescaling. A correction that repairs one property often damages another, so Tab. 3 scores every one we tried on all four at once.

The two constants. Two things are wrong, and they need different repairs. The level is off, so the average bin claims the wrong error by a factor, and the slope is below one, so the bins are wrong relative to each other. One number for the whole dataset fixes the level and nothing else, since it multiplies every bin alike and moves only the intercept of Eq. 5. A separate number per bin fixes both, but it damages the ranking. A power law fixes both and leaves the ranking alone. Eq. 5 fits $\ln \bar{e} = c + b \ln \hat{\sigma}$, so multiplying by

$$g(\hat{\sigma}) = e^c \hat{\sigma}^{b-1} \quad \text{gives} \quad \hat{\sigma}' = e^c \hat{\sigma}^b. \quad (6)$$

Writing $s = b - 1$ and $a = c + s \ln \alpha$ puts this in the form we release, $\hat{\sigma}' = \hat{\sigma} \exp(a + s \ln(\hat{\sigma}/\alpha))$, where a sets the level and $1 + s$ the slope. Over all 168 settings on held-out scenes it takes the median setting from $2.4\times$ off to $1.11\times$, and the slope from 0.65 to 0.99.

Ranking. A power law cannot change the ranking, since the map $\hat{\sigma} \mapsto e^c \hat{\sigma}^b$ is strictly increasing whenever the exponent is positive. The fitted exponent is positive on 162 of 168 settings, and negative but within 0.07 of zero on the rest. We clamp any exponent below 0.05, which affects ten settings. The within-image ρ is then identical to the native one, by measurement on the 158 settings we leave alone and by construction on the ten we clamp.

The same repair applied bin by bin does cost ranking, because a step correction can reorder pixels at a bin edge even where the smooth one cannot. It keeps the ranking

Correction <i>target</i>	par.	Ranking ρ		Level, delivered over predicted				Slope	Coverage at nominal 0.90	
		pooled	in-image	overall	low bin	high bin	spread	b	delivered	in band
		$\uparrow$	$\uparrow$	1	1	1	0	1	0.90	$\uparrow$
<i>Rescaling the predicted uncertainty</i>										
none, native	0	0.46	0.428	2.41 $\times$	1.48	3.60	0.258	0.65	0.570	0.000
one scalar per model	1	=	=	1.41 $\times$	0.56	1.33	0.258	=	0.917	0.133
one scalar per dataset [18, 33, 43]	1	=	=	1.14 $\times$	0.59	1.47	0.258	=	0.898	0.258
fitted on the other twelve	1	=	=	1.46 $\times$	0.55	1.37	0.258	=	0.914	0.106
one per confidence bin [5, 54]	14	*	0.376	1.11 $\times$	0.98	1.07	0.144	1.00	0.902	0.333
rank-preserving	≤ 14	*	0.428	1.14 $\times$	0.59	1.62	0.259	0.65	0.888	0.276
power law, two constants	2	=	=	1.11 $\times$	1.03	0.92	0.125	0.99	0.905	0.333
<i>Re-choosing the interval, leaving the predicted uncertainty untouched</i>										
isotonic [31]	1	=	=	=	=	=	=	=	0.893	0.267
four published refinements [‡]	7–70	=	=	=	=	=	=	=	—	+0.000
<i>Fitted on the scenes it is scored on, so a ceiling rather than a correction</i>										
one scalar per scene	1/scene	—	=	1.00 $\times$	0.69	1.42	0.198	0.72	0.914	0.867

Table 3. Every correction we tried, on all four properties, as grid medians over the 168 settings (VGGT under its default α , StreamVGGT against its teacher, Tab. 2), each fitted on a calibration half of a dataset’s scenes and scored on the other half, the per-model scalar on the backbone’s calibration halves pooled and “fitted on the other twelve” on those of the other datasets. The power-law row is scored before the exponent clamp of Sec. 6, after which its ranking cells are exact. par. counts the fitted constants and pooled ρ is on every scored pixel, as in Tab. 2. *target* is the value a column takes if the confidence is right, “=” is unchanged by construction and “—” is not scored. “spread” is how far the bins are off relative to each other, “delivered” the share of truth inside a nominal 0.90 interval and “in band” the fraction of held-out scenes within 0.05 of it, of which a setting holds a median of nine, from two to fifteen. Bold marks the deployable correction closest to a column’s target. The middle block re-chooses the interval and moves nothing to its left. *Paired change on 108 settings, median -0.001 . The per-scene scalar rescales scenes against each other, so its pooled ρ is not scored. [‡]Paired median change against the per-bin row.

only if its steps are non-increasing in the confidence, which the per-bin fit breaks wherever $b < 1$. Setting by setting, that costs a paired median 0.0138 of the within-image ρ .

Coverage. A held-out scene is in band when its own coverage lands within 0.05 of nominal, and a setting’s figure is the fraction of its held-out scenes that do. At nominal 0.90, one scalar for the whole model puts 13 % of a setting’s scenes in band, one per dataset 26 %, and one per bin 33 %. Read per clip and per frame instead of per scene, every one of these corrections lands fewer units in band (supplement, Fig. S3). Nothing we tried beyond one scalar per bin does better, and in the classification case distribution-free calibration is known to be reachable only bin by bin [19]. Coverage therefore cannot separate the corrections at or beyond one scalar per bin, and the choice depends on what else the number is for.

A scalar per dataset is enough if all that matters is an interval that holds its stated rate. The power law is the one to use if the uncertainty itself has to be right at every confidence level, since it is the only correction here that repairs the slope without costing ranking. A bundle adjustment weighting residuals by $\hat{\sigma}^{-2}$ is one such use [25], since a slope below one over-weights the pixels the model is most confident about.

Released constants. Tab. 4 gives the fitted pair for every backbone and dataset we ran, and a median row to borrow for a dataset not among them. Scored on a dataset held out of the median that produced it, it takes the average confidence bin from 2.41 $\times$ off to 1.35 $\times$ and helps on 74 % of settings. A pair refitted on labelled scenes of the target dataset reaches 1.12 $\times$ and helps on 91 % (Tab. 3’s 1.11 $\times$ is the same fit before the exponent clamp). Holding out by corpus rather than by dataset name changes nothing. The borrowed pair makes about a quarter of settings worse, and Tab. 4 shows which backbones.

Scene to scene. A correction blind to the scene multiplies every scene alike, so it cannot change how far apart two scenes are. Every correction above is close to right on average and still lands at most a third of held-out scenes in band. A scalar fitted on each held-out scene’s own labels reaches 0.867, a paired median gain of 0.42 [0.33, 0.47] that is positive on 148 of 168 settings (supplement, Fig. S3). Both figures are fractions of a median of nine held-out scenes per setting, from two on VKITTI2 to fifteen on ScanNet and DL3DV, so neither is finer than about a ninth. That scalar is fitted on the scenes it is scored on, so it measures the gap rather than closing it.

The same scalar read per clip scores 0.333, and per frame 0.212 (supplement, Fig. S3). Fitting one scalar per

Dataset	DUST3R		MonST3R		MAST3R		CUT3R		St4RTrack		StreamVGGT		VGGT	
	a	$1+s$	a	$1+s$	a	$1+s$	a	$1+s$	a	$1+s$	a	$1+s$	a	$1+s$
ScanNet	-0.36	0.70	+0.21	0.81	-0.09	0.38	-0.12	0.35	+0.07	0.32	-0.26	0.90	-0.82	0.63
ScanNet-tr	-0.42	0.72	+0.22	0.77	-0.13	0.35	+0.04	0.45	+0.15	0.37	-0.18	0.94	-0.95	0.62
7-Scenes	+0.12	0.72	+0.75	0.84	+0.24	0.38	+0.38	0.57	+0.59	0.23	-0.09	0.92	-0.24	0.45
TUM	+0.06	0.78	+0.59	0.81	+0.29	0.19	+0.59	0.47	+0.88	0.56	-0.25	0.93	-0.09	0.52
TUM-dyn	+0.73	1.03	+0.94	0.97	+0.76	0.65	+1.12	1.08	+0.46	0.73	+0.28	1.09	-0.37	0.52 [†]
Bonn	-0.07	0.41	+0.39	0.55	+0.16	0.34	+0.54	0.61	+1.30	0.42	-0.03	1.01	-0.31	0.37
DL3DV	+0.38	0.51	+0.48	0.65	+1.44	0.34	+1.60	0.41	+1.55	0.42	+0.04	1.04	+0.96	0.57
CO3Dv2	+0.12	0.86	+0.09	0.77	+1.41	0.05 [†]	+1.27	0.05 [†]	-	-	+0.25	1.01	+0.40	0.39
BlendedMVS	+0.44	0.69	+0.47	0.84	+1.33	0.07 [†]	+1.50	0.21 [†]	+1.03	0.55	+0.19	1.06	-1.03	0.08 [†]
KITTI	+0.79	0.91	+0.73	0.89	+1.46	0.40	+0.56	1.07	+0.72	0.05 [†]	+0.21	1.02	+0.44	0.74
VKITTI2	+0.70	0.83	+0.61	0.68	+1.60	0.50	+0.01	0.93	+1.16	0.06	+0.18	1.07	-0.42	0.48
Sintel	+1.02	0.77	+0.50	0.74	+0.79	0.54	+0.77	0.96	+1.05	0.05 [†]	-0.02	0.92	+0.51	0.68
Dyn.Replica	-0.07	0.85	+0.28	0.62	-0.31	0.61	+0.02	0.63	+0.41	0.70	-0.28	1.01	-0.81	0.59
<i>median</i>	+0.19	0.78	+0.42	0.79	+0.70	0.38	+0.62	0.49	+0.80	0.39	-0.04	0.98	-0.22	0.51
level, native	1.78		2.54		5.65		3.39		3.00		1.11		1.90	
with the median pair	1.42		1.24		2.22		1.74		1.58		1.13		1.52	
with a refit pair	1.12		1.14		1.18		1.13		1.06		1.06		1.12	

Table 4. The released correction, one entry per backbone and dataset. Each entry is a and the exponent $1+s$ of Eq. 6, the median over that entry’s heads, fitted on a calibration half of its scenes. Apply as $\hat{\sigma}' = \hat{\sigma} \exp(a + s \ln(\hat{\sigma}/\alpha))$. The *median* row is over that backbone’s settings, not over the column. The three rows below score it on held-out scenes, “with the median pair” on a dataset held out of the median that produced it and “with a refit pair” on a few labelled scenes of the target, as how far the average bin is off. A dagger marks a pair with at least one head’s exponent clamped at 0.05.

clip or per frame does not rescue it either, since a correction handed the right number for every frame still leaves three frames in five outside the band.

What closes it on a scene is that scene’s own empirical quantiles at the eleven levels rather than any one number for it. Fitted on the scene, they put every held-out scene in band by construction, but read per clip and per frame only 0.41 and 0.26 of units. So what the confidence is missing is a shape rather than a shift, and it varies faster than any unit we can name.

Estimating it without labels is the open question, and we could not answer it. Ground truth on an opening window of each scene, applied to the rest, never beats a correction that ignores the scene. Labels from *inside* the corrected window do help, but the gain survives shuffling the frames into a random order, so it is averaging noise rather than following anything.

Conformal baselines. Four published methods condition on more than the bin alone, by a richer partition of frames [54], by overlapping groups [24], by distance to the test frame [17], or by a density ratio between calibration and deployment [51]. Run against the same per-bin baseline on all 168 settings, none moves the median off zero (Tab. 3). Each improves 15 to 36 % of settings and damages a similar share, and adaptive conformal prediction [15] does no better when given ground truth on an opening window of up to half the frames.

The last of the four is the informative one, because the shift it corrects is really there. A classifier separates calibration frames from held-out frames at AUC 0.68 using the predictions alone, yet correcting for that shift leaves as

many settings worse as better. What differs between calibration and deployment is not which frames are drawn but what the error does at a given confidence.

None of this is ruled out in advance. Exact coverage cannot be promised at a single point in feature space [4, 34, 54], but a scene is a group of 10^5 to 10^6 pixels rather than a point. Whether what is missing varies smoothly with something observable about a scene we tested and could not settle, with too few scenes per setting to tell a real effect from none.

Outside the class. Two repairs sit outside it. A linear probe on the model’s frozen features, rather than on its confidence value alone, closes a median 0.94 of the slope gap out of sample on the six of nine settings of three backbones that pass a validity screen. What transfers there is the level and the slope, not the ranking. Expressing the prediction in the frame’s own predicted scale needs no labels and recovers half of the remaining coverage gap on MAST3R’s metric branch ($n = 14$), against a transductive ceiling and leaving the non-metric branch where it was. Improving the ranking needs a different head on the same features [62]. The protocol, the per-setting outputs and the fitted constants accompany this submission.

7. Conclusion

Native confidence is a usable ranking signal with an unreliable level. It ranks error inside images almost everywhere and understates that error, the more so the more confident the model is. Post-hoc repair recovers the level at no cost to the ranking, once the correction is written as a power law rather than a step function, and not much further. We release the protocol and the fitted constants.

References

- [1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 14927–14937, 2020.
- [2] Anastasios N. Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael I. Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *Proceedings of the 39th International Conference on Machine Learning*, pages 717–730. PMLR, 2022.
- [3] Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.
- [4] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.
- [5] Henrik Boström and Ulf Johansson. Mondrian conformal regressors. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, pages 114–133. PMLR, 2020.
- [6] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 611–625. Springer, 2012.
- [7] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. arXiv:2001.10773 [cs.CV], 2020.
- [8] Liyan Chen, Weihang Wang, and Philippos Mordohai. Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17235–17244, 2023.
- [9] João Marcos Correia Marques, Albert J. Zhai, Shenlong Wang, and Kris Hauser. On the overconfidence problem in semantic 3D mapping. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 11095–11102, 2024.
- [10] Jan Czarnowski, Tristan Laidlow, Ronald Clark, and Andrew J. Davison. DeepFactors: Real-time probabilistic dense monocular SLAM. *IEEE Robotics and Automation Letters*, 5(2):721–728, 2020.
- [11] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017.
- [12] Haiwen Feng, Junyi Zhang, Qianqian Wang, Yufei Ye, Pengcheng Yu, Michael J. Black, Trevor Darrell, and Angjoo Kanazawa. St4RTrack: Simultaneous 4D reconstruction and tracking in the world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8503–8513, 2025.
- [13] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016.
- [14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.
- [15] Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1660–1672, 2021.
- [16] Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes’ Rays: Uncertainty quantification for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20061–20070, 2024.
- [17] Laying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.
- [18] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.
- [19] Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: Prediction sets, confidence intervals and calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3711–3723, 2020.
- [20] Markus Hillemann, Robert Langendörfer, Steven Landgraf, and Markus Ulrich. Uncertainty quality of VGGT: An analysis on the DTU benchmark dataset. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XI-2-2026:665–672, 2026.
- [21] Julia Hornauer and Vasileios Belagiannis. Gradient-based uncertainty for monocular depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 613–630. Springer, 2022.
- [22] Eddy Ilg, Özgün Çiçek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 677–693. Springer, 2018.
- [23] Seonghyun Jin and Jong Chul Ye. FILT3R: Latent state adaptive Kalman filter for streaming 3D reconstruction. *arXiv preprint arXiv:2603.18493*, 2026.
- [24] Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2023.
- [25] Jaehyung Jung, Simon Boche, Sebastián Barbas Laina, and Stefan Leutenegger. Uncertainty-aware visual-inertial SLAM with volumetric occupancy mapping. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.
- [26] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent dynamic depth from stereo videos.

- In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13229–13239, 2023.
- [27] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9492–9502, 2024.
 - [28] Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. Early access.
 - [29] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5574–5584, 2017.
 - [30] Simon Kristoffersson Lind, Ziliang Xiong, Per-Erik Forssén, and Volker Krüger. Uncertainty quantification metrics for deep regression. *Pattern Recognition Letters*, 186:91–97, 2024.
 - [31] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2796–2804. PMLR, 2018.
 - [32] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6402–6413, 2017.
 - [33] Max-Heinrich Laves, Sontje Ihler, Jacob F. Fast, Lüder A. Kahrs, and Tobias Ortmaier. Recalibration of aleatoric and epistemic regression uncertainty in medical imaging. *Machine Learning for Biomedical Imaging*, 1, 2021.
 - [34] Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B*, 76(1):71–96, 2014.
 - [35] Jing Lei, Max G’Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.
 - [36] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MAST3R. In *Computer Vision – ECCV 2024*, pages 71–91. Springer, 2024.
 - [37] Dan Levi, Liran Gispán, Niv Giladi, and Ethan Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors*, 22(15):5540, 2022.
 - [38] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22160–22169, 2024.
 - [39] Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. SLAM3R: Real-time dense scene reconstruction from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16651–16662, 2025.
 - [40] Rémi Marsal, Florian Chabot, Angélique Loesch, William Grolleau, and Hichem Sahbi. MonoProb: Self-supervised monocular depth estimation with interpretable uncertainty. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3637–3646, 2024.
 - [41] David A. Nix and Andreas S. Weigend. Estimating the mean and variance of the target probability distribution. In *IEEE International Conference on Neural Networks (ICNN)*, pages 55–60, 1994.
 - [42] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguère, and Cyrill Stachniss. ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7855–7862, 2019.
 - [43] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, pages 345–356. Springer, 2002.
 - [44] 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*, 48(3):2354–2367, 2026.
 - [45] John C. Platt. Probabilities for SV machines. In *Advances in Large-Margin Classifiers*, pages 61–74. MIT Press, 2000.
 - [46] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3227–3237, 2020.
 - [47] Matteo Poggi, Seungryong Kim, Fabio Tosi, Sunok Kim, Filippo Aleotti, Dongbo Min, Kwanghoon Sohn, and Stefano Mattoccia. On the confidence of stereo matching in a deep-learning era: A quantitative evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5293–5313, 2022.
 - [48] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10901–10911, 2021.
 - [49] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2930–2937, 2013.
 - [50] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the eval-

- uation of RGB-D SLAM systems. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 573–580, 2012.
- [51] Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
 - [52] Stephanie Tsuei, Stefano Soatto, Paulo Tabuada, and Mark B. Milam. Learned uncertainty calibration for visual inertial localization. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5311–5317, 2021.
 - [53] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant CNNs. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 11–20, 2017.
 - [54] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, pages 475–490. PMLR, 2012.
 - [55] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5306, 2025.
 - [56] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10510–10522, 2025.
 - [57] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUST3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024.
 - [58] Heng Yang and Marco Pavone. Object pose estimation with statistical guarantees: Conformal keypoint detection and geometric uncertainty propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8947–8958, 2023.
 - [59] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1790–1799, 2020.
 - [60] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *The Thirteenth International Conference on Learning Representations*, 2025.
 - [61] Wei Zhang, Yihang Wu, Songhua Li, Wenjie Ma, Xin Ma, Qiang Li, and Qi Wang. Review of feed-forward 3D reconstruction: From DUST3R to VGGT. *arXiv preprint arXiv:2507.08448*, 2025.
 - [62] Zihao Zhu, Wenyuan Zhao, Nuo Chen, Chao Tian, and Zhiwen Fan. Trust it or not: Evidential uncertainty for feed-forward 3D reconstruction with Trust3R. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.