RANSAC Scoring Done Right

James Pritts **Felix Seegräber** **Kevin Köser**
Marine Data Science, Institut für Informatik
Kiel University, Germany
{jpr, fse, kk}@informatik.uni-kiel.de

Abstract

The most widely used RANSAC variants score a candidate model by summing per-point scores that fall to zero beyond a threshold. These scores derive from the negative log-likelihood of a Gaussian–uniform mixture. The threshold separates inliers, which follow the noise model, from outliers, and it is a function of the observation noise level, also known as the inlier scale. We remove this dependence on the threshold by reversing the usual order of RANSAC inference: rather than estimating the scale and then scoring against it, we marginalize the scale in closed form for a fixed inlier partition and then optimize over partitions. The resulting score is maximized by a single sort followed by a linear sweep. A single closed-form expression spans informative and non-informative priors, so the same score adapts to datasets with extensive or limited ground-truth annotation. The proposed score has a single hyperparameter, the half-width of the uniform outlier distribution. We show that this half-width is far less sensitive to miscalibration than the inlier threshold and is easier to cross-validate. The proposed score is the first likelihood-based RANSAC score in which the inlier scale is absent from the formula.[1]

1 Introduction

RANSAC [14] scores candidate geometric models by evaluating how well they explain the observed data in the presence of outliers. Every widely used scoring function, whether inlier counting (RANSAC), truncated squared error (MSAC [35]), or the Gaussian–uniform marginal likelihood (MLESAC [37]), depends on a hyperparameter that implicitly encodes the inlier scale. The user must either supply this quantity or estimate it from contaminated data, neither of which is straightforward in practice.

Shekhovtsov [33] provides a unifying analysis of these functions through the lens of a single probabilistic model: the Gaussian–uniform mixture, in which each inlier residual follows a Gaussian density and each outlier is uniform. Within this framework, MSAC is the *profile* likelihood (maximizing over latent inlier labels) and MLESAC is the *marginal* likelihood (summing over labels), the score Shekhovtsov calls GaU. Both are parameterized by a decision threshold $\tau = q\,\sigma$, a fixed quantile q of the inlier scale σ, so tuning the threshold is equivalent to tuning the inlier scale in different units. We call scores of this form *threshold-based*.

We propose an *inlier-scale-free* RANSAC score, the *profile-marginal score*, by reversing the usual order of elimination. Threshold-based scores sum or maximize over the latent inlier/outlier labels first, retaining the inlier scale through the threshold. We instead integrate out the inlier scale σ for each fixed partition and then maximize over partitions. To our knowledge, this is the first likelihood-based RANSAC score to eliminate σ in this order. A conjugate Inverse-Gamma prior on σ^2 gives a closed form that also covers the Jeffreys limit. When annotated validation data are available, we estimate the prior parameters by empirical Bayes. Otherwise, we use the Jeffreys limit with the same scoring algorithm.

[1] Project page, with code and the reproduction project: `https://ransacscoringdoneright.github.io`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

For a fixed prior, the score depends on the inlier residual sum of squares, the inlier count, and the half-width a of the outlier domain (Section 3). We maximize it by sorting the point-to-model distances; the best inlier set is always a prefix of the sorted list, so one sweep over the prefixes finds it. This takes $O(N)$ time with the radix sort used in our implementation and $O(N \log N)$ with a comparison sort. The half-width a parameterizes the outlier model and is not tied to any quantile of the inlier scale. When an image pair contains enough inliers for the data to dominate the prior, we find that the score is significantly less sensitive to a than threshold-based scores are to their threshold. The score is thus inlier-scale-free, and Section D compares our score with Torr's Bayesian framework and GRIC [36].

MAGSAC [4] and MAGSAC++ [5] were introduced specifically to remove the inlier-scale dependence by "marginalizing" the inlier scale under the Gaussian–uniform assumption. Shekhovtsov [33] shows that the derivation's errors cancel by numerical coincidence, leaving a score numerically equivalent to that of MLESAC with a fixed inlier scale. The MAGSAC++ hyperparameter is defined as the upper bound on the inlier scale, but it is in fact a proxy for the mixture's inlier scale (Section D). Moreover, when the threshold is properly tuned, *all* of these scoring functions (MSAC, GaU, MAGSAC++) perform nearly identically for model selection. Only inlier counting (standard RANSAC) is measurably worse. The bottleneck is not the choice of scoring function but the dependence on a well-chosen inlier-scale quantile.

Other work avoids a hand-set threshold. Adaptive-scale estimators (ASSC [38], ASKC [39]) and SIMFIT++ [12] estimate the scale from the contaminated distances and then score against it. This is the estimate-then-score order we reverse. RECON [28] instead seeks consistency in how different models rank the points by distance. A-contrario methods (MINPRAN [34], ORSA/AC-RANSAC [25, 26]) are closest in spirit: they also sweep distance-sorted prefixes and require only an outlier domain, but they rank inlier sets by a significance test against a uniform background model rather than by a likelihood. Our score is the marginal likelihood of the same Gaussian–uniform model that the threshold-based scores share.

2 Probabilistic Model for Robust Fitting

Consider N observations $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, \dots, N$, and a candidate geometric model $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$. The noise-free point $\bar{\mathbf{x}}_i$ of each inlier lies on the model manifold $\mathcal{M}_{\boldsymbol{\theta}} = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{g}(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{0}\}$, whose codimension $d_g = d - \dim \mathcal{M}_{\boldsymbol{\theta}}$ is the number of directions normal to it. Noise along $\mathcal{M}_{\boldsymbol{\theta}}$ leaves $\mathbf{g}$ unchanged to first order, so only the d_g normal directions carry evidence. The codimension therefore fixes the dimension of the residual vector, and each observation imposes d_g independent scalar constraints on $\boldsymbol{\theta}$, one per normal direction. A minimal sample has $m = \lceil n_\theta / d_g \rceil$ observations, the fewest whose constraints determine $\boldsymbol{\theta}$ up to finitely many solutions. A minimal solver computes these solutions in the hypothesis step that proposes models for scoring; Table 1 lists (n_θ, d_g, m) for common geometric models.

An inlier is a noisy observation of a point on the model, $\mathbf{x}_i = \bar{\mathbf{x}}_i + \boldsymbol{\varepsilon}_i$ with $\mathbf{g}(\bar{\mathbf{x}}_i, \boldsymbol{\theta}) = \mathbf{0}$ and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}_i})$, where the shape $\widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}_i}$ is known and the inlier scale σ is not. To first order $\mathbf{g}_i \coloneqq \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}) \approx \mathbf{J}_i \boldsymbol{\varepsilon}_i$, with $\mathbf{J}_i$ the Jacobian of $\mathbf{g}$ with respect to $\mathbf{x}$. Whitening by $\mathbf{L}_i \mathbf{L}_i^\top = \mathbf{J}_i \widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}_i} \mathbf{J}_i^\top$ gives the residual

$$\mathbf{r}_i \coloneqq \mathbf{L}_i^{-1} \mathbf{g}_i \;\sim\; \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{d_g}), \tag{1}$$

whose squared norm $\mathbf{g}_i^\top (\mathbf{J}_i \widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}_i} \mathbf{J}_i^\top)^{-1} \mathbf{g}_i$ is the general form of the squared Sampson distance [30, 17, 7]. Equivalently, it is the squared Mahalanobis distance, with respect to $\widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}_i}$, from $\mathbf{x}_i$ to the linearization of $\mathcal{M}_{\boldsymbol{\theta}}$ at $\mathbf{x}_i$. Its relation to the geometric distance to $\mathcal{M}_{\boldsymbol{\theta}}$ is discussed in [16, 29]. Our model requires only that the distribution of the residual be isotropic normal.

We adopt the standard Gaussian–uniform mixture [37, 33]: a latent label $z_i \in \{0, 1\}$ marks observation i as inlier (z_i=1) or outlier (z_i=0) with prior $\Pr(z_i{=}1) = \gamma$, and the residual $\mathbf{r}_i$ has conditional density

$$p(\mathbf{r}_i \mid z_i, \boldsymbol{\theta}, \sigma) = \begin{cases} p_{\text{in}}(\mathbf{r}_i; \sigma) \coloneqq (2\pi\sigma^2)^{-d_g/2} \exp\!\big(-\|\mathbf{r}_i\|^2 / 2\sigma^2\big) & z_i = 1, \\ p_{\text{out}} \coloneqq (2a)^{-d_g} & z_i = 0, \end{cases} \tag{2}$$

a Gaussian inlier density and a uniform outlier density on $[-a, a]^{d_g}$; the half-width a has the same units as σ (e.g. pixels for a Sampson distance). We normalize coordinates so that the image half-

diagonal is one, which fixes the outlier domain, and hence a, independently of σ. By independence, the joint density of residuals and labels $\mathbf{z} = (z_1, \ldots, z_N)$ factorizes as

$$p(\mathbf{r}, \mathbf{z} \mid \boldsymbol{\theta}, \sigma) = \prod_{i=1}^{N} \big[\gamma\, p_{\text{in}}(\mathbf{r}_i; \sigma)\big]^{z_i} \, \big[(1-\gamma)\, p_{\text{out}}\big]^{1-z_i}. \tag{3}$$

Maximizing (3) over labels classifies $z_i^* = 1$ iff $\gamma\, p_{\text{in}}(\mathbf{r}_i; \sigma) > (1-\gamma)\, p_{\text{out}}$; collecting positive labels defines the inlier set $\mathcal{I} \coloneqq \{i : z_i^* = 1\}$ with complement $\mathcal{O} = \{1, \ldots, N\} \setminus \mathcal{I}$, and $n_I = |\mathcal{I}|$, $n_O = |\mathcal{O}|$. This per-point rule is the profile-likelihood labeling at fixed σ; in Section 3 we instead treat the partition $\mathcal{I}$ as a free variable to optimize. Given a partition, the likelihood factorizes as

$$p(\mathbf{r} \mid \boldsymbol{\theta}, \mathcal{I}, \sigma) = \prod_{i \in \mathcal{I}} \frac{1}{(2\pi\sigma^2)^{d_g/2}} \, \exp\Big(-\frac{\|\mathbf{r}_i\|^2}{2\sigma^2}\Big) \;\cdot\; \frac{1}{(2a)^{n_O d_g}}\,. \tag{4}$$

3 Profile-Marginal Likelihood Scoring

Besides the model $\boldsymbol{\theta}$, the joint (3) has two nuisance quantities, the labels $\mathbf{z}$ and the scale σ. Scoring a candidate requires eliminating both. Standard approaches eliminate the labels first—by summation (MLESAC's marginal likelihood [37]) or maximization (MSAC's profile likelihood [35])—and the resulting scores still depend on σ [33].

3.1 Reversing the Elimination Order

We take a different path. Rewrite the joint (3) by grouping labels into a partition $\mathcal{I} = \{i : z_i = 1\}$:

$$p(\mathbf{r}, \mathbf{z} \mid \boldsymbol{\theta}, \sigma) = \gamma^{n_I}(1-\gamma)^{n_O} \;\cdot\; p(\mathbf{r} \mid \boldsymbol{\theta}, \mathcal{I}, \sigma), \tag{5}$$

where $p(\mathbf{r} \mid \boldsymbol{\theta}, \mathcal{I}, \sigma)$ is the partition-conditioned likelihood (4) and the label prior $\gamma^{n_I}(1-\gamma)^{n_O}$ depends only on $n_I = |\mathcal{I}|$. For general γ the label prior adds $n_I \log\frac{\gamma}{1-\gamma}$ (plus a constant) to the log-score, biasing the inlier/outlier boundary. We adopt the non-informative choice $\gamma = \frac{1}{2}$, for which the label prior is the constant 2^{-N} across all partitions and drops out of the optimization.[2]

Rather than eliminating labels first (which entangles σ with the per-point sums or maxima), we *reverse the order*: for a *fixed* partition $\mathcal{I}$ and a fixed candidate $\boldsymbol{\theta}$, a plug-in for the minimal-sample solution as in all RANSAC scores (Section A.1), we marginalize σ under a conjugate Inverse-Gamma prior $\pi(\sigma^2) = \text{IG}(\alpha_0, \beta_0)$, defining the *scale-marginalized log-likelihood*

$$S(\boldsymbol{\theta}, \mathcal{I}) \;\coloneqq\; \log \int_0^\infty p(\mathbf{r} \mid \boldsymbol{\theta}, \mathcal{I}, \sigma)\; \pi(\sigma^2)\, d\sigma^2, \tag{6}$$

and then optimize over partitions, giving the model score

$$S^*(\boldsymbol{\theta}) \coloneqq \max_{\mathcal{I}} \; S(\boldsymbol{\theta}, \mathcal{I}). \tag{7}$$

We call S^* the *profile-marginal score*: profile over labels $\mathbf{z}$ (equivalently partitions $\mathcal{I}$), marginal over σ. A fully Bayesian score would marginalize *both*, but $\log \sum_{\mathcal{I}} \exp S(\boldsymbol{\theta}, \mathcal{I})$ has 2^N terms; the hybrid avoids this sum while still eliminating σ analytically, because for a fixed partition the σ-integral (6) has a closed form (Section 3.2).

3.2 The Profile-Marginal Score in Closed Form

The Inverse-Gamma is the conjugate prior for the variance of a zero-mean Gaussian, so (6) has a closed form for any $\alpha_0, \beta_0 > 0$; we keep (α_0, β_0) general, so the score holds for every prior in the family. Taking logarithms yields the closed form of S, from which the profile-marginal score (7) follows by maximizing over $\mathcal{I}$ (Section A gives the full derivation):

$$\boxed{S(\boldsymbol{\theta}, \mathcal{I}) = \log\Gamma\Big(\frac{\nu}{2} + \alpha_0\Big) - \Big(\frac{\nu}{2} + \alpha_0\Big)\log\Big(\frac{\text{RSS}_{\mathcal{I}}}{2} + \beta_0\Big) - \frac{\nu}{2}\log(2\pi) - n_O\, d_g\, \log(2a),} \tag{8}$$

[2] One could instead marginalize γ under a conjugate Beta prior, exactly as we marginalize σ: $\int_0^1 \gamma^{n_I}(1-\gamma)^{n_O}\, \pi(\gamma)\, d\gamma$ is closed-form and, depending only on the count n_I, preserves the sort-and-sweep optimality and its cost. We fix $\gamma = \frac{1}{2}$ instead so that the label prior is partition-independent and the consensus decision is driven solely by fit quality and outlier volume rather than a separately estimated inlier rate, which itself depends on the partition.

where $\mathrm{RSS}_{\mathcal{I}} = \sum_{i\in\mathcal{I}} \|\mathbf{r}_i\|^2$ is the inlier residual sum of squares and $\nu = n_I\, d_g - n_\theta$ its degrees of freedom: the $n_I\, d_g$ Gaussian residual coordinates of the inliers less the n_θ consumed by fitting $\boldsymbol{\theta}$ (Section A.1 derives this correction).

Pseudo-observation reading. The integrand of (6) is proportional to an Inverse-Gamma density in σ^2, so the posterior given the inlier residuals is

$$\sigma^2 \mid \mathcal{I} \;\sim\; \mathrm{IG}\Big(\tfrac{2\alpha_0+\nu}{2},\; \tfrac{2\beta_0+\mathrm{RSS}_{\mathcal{I}}}{2}\Big). \tag{9}$$

Each inlier i outside the minimal sample contributes d_g residual coordinates to ν and $\|\mathbf{r}_i\|^2$ to $\mathrm{RSS}_{\mathcal{I}}$ (so $\nu = (n_I - m)\, d_g$ for an exact minimal solver, Section A.1). Without prior information (the Jeffreys limit) the posterior is $\mathrm{IG}(\nu/2, \mathrm{RSS}_{\mathcal{I}}/2)$; comparing with (9), the prior only replaces the coordinate count ν by $\nu + 2\alpha_0$ and the sum of squares $\mathrm{RSS}_{\mathcal{I}}$ by $\mathrm{RSS}_{\mathcal{I}} + 2\beta_0$. It thus enters the posterior exactly as $2\alpha_0$ additional residual coordinates would, the *pseudo-observations*, with sum of squares (*pseudo-RSS*) $2\beta_0$. Their mean square $s_0^2 = 2\beta_0/2\alpha_0 = \beta_0/\alpha_0$ is the prior's guess of σ^2, just as $\mathrm{RSS}_{\mathcal{I}}/\nu$ is the data's. Whether fitted by empirical Bayes (Section C gives two estimators) or chosen by hand, the prior thus specifies both s_0^2 and the number $2\alpha_0$ of coordinates supporting it. As a point estimate of the inlier variance we take the mean square of the real and pseudo-observations together, which reduces to the classical $\mathrm{RSS}_{\mathcal{I}}/\nu$ in the Jeffreys limit (and equals $1/\mathbb{E}[\sigma^{-2} \mid \mathcal{I}]$, Section A); for $\nu > 0$,

$$\hat{s}^2 \coloneqq \frac{2\beta_0 + \mathrm{RSS}_{\mathcal{I}}}{2\alpha_0 + \nu} = \frac{2\alpha_0}{2\alpha_0 + \nu}\, s_0^2 + \frac{\nu}{2\alpha_0 + \nu}\, \frac{\mathrm{RSS}_{\mathcal{I}}}{\nu}.$$

Marginally, an inlier residual coordinate is then Student-t with $2\alpha_0 + \nu$ degrees of freedom and scale $\hat{s}$. The prior mean square dominates when $\nu \ll 2\alpha_0$ and the observed one when $\nu \gg 2\alpha_0$, which is how the prior regularizes the variance estimate in the data-scarce regime of Section 4.

The four terms of the score. All four terms of (8) depend on the partition $\mathcal{I}$:

1. *Inlier count.* $\log\Gamma(\nu/2 + \alpha_0)$ grows super-linearly for large ν, favoring larger consensus sets.
2. *Fit quality.* $-(\nu/2 + \alpha_0)\log(\mathrm{RSS}_{\mathcal{I}}/2 + \beta_0)$ favors smaller residual sums of squares at fixed ν. A positive pseudo-RSS $2\beta_0$ keeps the term finite as $\mathrm{RSS}_{\mathcal{I}} \to 0$.
3. *Gaussian normalizer.* $-(\nu/2)\log(2\pi)$ supplies the normalization constant for the ν inlier coordinates.
4. *Outlier penalty.* $-n_O\, d_g \log(2a)$ assigns each outlier the uniform-model cost $d_g \log(2a)$.

Non-informative limit (Jeffreys prior). The limit $(\alpha_0, \beta_0) \to (0, 0)$ gives the improper Jeffreys prior $\pi(\sigma^2) \propto 1/\sigma^2$ [18], which introduces no scale assumption. Its posterior is proper whenever $\nu > 0$, so we restrict the maximization in (7) to partitions with $\nu > 0$, i.e. $n_I > n_\theta/d_g$. For example, with an exact minimal solver, at least one inlier beyond the minimal sample is required. The score (8) then simplifies to

$$S_J(\boldsymbol{\theta}, \mathcal{I}) = \log\Gamma\Big(\frac{\nu}{2}\Big) - \frac{\nu}{2}\,\log(\pi\ \mathrm{RSS}_{\mathcal{I}}) - n_O\, d_g\ \log(2a), \tag{10}$$

where the combination $(\nu/2)\log(2\pi) + (\nu/2)\log(\mathrm{RSS}_{\mathcal{I}}/2) = (\nu/2)\log(\pi\ \mathrm{RSS}_{\mathcal{I}})$ has been merged. Every term is data-only, without any pseudo-observation contribution from the prior, so S_J is the natural baseline against which an informative prior adds value only when ν is small, i.e. when there are too few inlier residuals to constrain σ (the weighted average above).

Implicit threshold. Although the score eliminates σ, the boundary of the optimal partition $\mathcal{I}^*$ defines an *implicit* threshold that, in the large-sample limit, satisfies the MAP condition $p_{\mathrm{in}}(\mathbf{r}; \hat{s}) = p_{\mathrm{out}}$ at $\|\mathbf{r}\| = \tau^*$ (the inlier density depends on $\mathbf{r}$ only through $\|\mathbf{r}\|$), with $\hat{s}$ the variance estimate above evaluated at $\mathcal{I}^*$ ($\hat{s}^2 = \mathrm{RSS}_{\mathcal{I}^*}/\nu$ in the Jeffreys limit):

$$\tau^* = \hat{s}\,\sqrt{2\, d_g \log\big(a\sqrt{2/\pi}\,/\,\hat{s}\big)}, \qquad \frac{\partial \log \tau^*}{\partial \log a} = \frac{1}{2\log\big(a\sqrt{2/\pi}\,/\,\hat{s}\big)}. \tag{11}$$

Unlike the baselines' one global τ, the proposed profile-marginal score adapts per hypothesis through $\hat{s}$ (Section B derives it). It also explains the robustness to miscalibration: a baseline's boundary is proportional to its hyperparameter, whereas τ^* depends on a only through a square-rooted logarithm: by (11), for any d_g, a 1% change in a changes τ^* by only 0.1–0.25%, and even overestimating a tenfold widens τ^* by only 24–45%. The threshold is the MAP threshold of the Gaussian–uniform mixture at $\gamma = \frac{1}{2}$, and for $a/\hat{s} \in [10, 100]$ it gives $\tau^* \approx (2\text{–}3)\sqrt{d_g}\,\hat{s}$, recovering the familiar χ^2-quantile rules of thumb (1.96σ–3σ for $d_g = 1$, 2.45σ–3.4σ for $d_g = 2$ [17]) as a Bayes decision boundary, which balances rejected inliers against accepted outliers, rather than one that fixes only the rate of rejected inliers: the boundary sits where the Gaussian and the outlier density are equally likely, and tightens automatically as outliers become denser.

3.3 The Sort-and-Sweep Algorithm

For a fixed $\boldsymbol{\theta}$, the distances $\delta_i = \|\mathbf{r}_i\|$ depend only on $\boldsymbol{\theta}$, not on the partition. Sorting them $\delta_{(1)} \le \cdots \le \delta_{(N)}$, the optimal inlier set is always a prefix $\mathcal{I}_k = \{(1), \ldots, (k)\}$ of some length k, written $\mathcal{I}^* = \mathcal{I}_{k^*}$ at the optimum: swapping any included point for an excluded one with a smaller distance lowers $\mathrm{RSS}_{\mathcal{I}}$ at fixed n_I and strictly raises the score. Since $\mathrm{RSS}_{\mathcal{I}_k} = \sum_{j=1}^{k} \delta_{(j)}^2$ is then a prefix sum, a running accumulator evaluates the full score in one $O(N)$ pass after the sort, starting at $k = \lfloor n_\theta / d_g \rfloor + 1$, the smallest length with $\nu = k\,d_g - n_\theta > 0$ (Section 3.2); the Jeffreys limit ($\alpha_0{=}\beta_0{=}0$) is the same algorithm without the pseudo-observations. Algorithm 1 summarizes the procedure. Within RANSAC, the sweep is applied to each candidate model generated from a minimal sample [14]; the model with the highest score S^* is selected. The sort sets the cost: $O(N \log N)$ with a comparison sort, and $O(N)$ with a radix sort on fixed-precision keys (we sort 32-bit floats, the precision of the residuals), the same order as MSAC at 9–20 μs more per hypothesis (Table 2).

Algorithm 1 Sort-and-sweep computation of the profile-marginal score (7), using (8). The algorithm returns the best-scoring prefix $\mathcal{I}^* = \{(1), \ldots, (k^*)\}$ with one sort and an $O(N)$ sweep.

Require: distances $\delta_1, \ldots, \delta_N$; codimension d_g; model parameters n_θ; outlier half-width a; prior parameters (α_0, β_0)
Ensure: best score S^*; optimal inlier count k^*

Sort so that $\delta_{(1)} \le \delta_{(2)} \le \cdots \le \delta_{(N)}$
$k_{\min} \leftarrow \min\big(\lfloor n_\theta / d_g \rfloor,\, N\big)$ ▷ largest prefix with $\nu \le 0$
$\mathrm{RSS} \leftarrow \sum_{j=1}^{k_{\min}} \delta_{(j)}^2$ ▷ seed the zero-DOF prefix
$S^* \leftarrow -\infty, \quad k^* \leftarrow 0$
for $k = k_{\min} + 1, \ldots, N$ **do**
  $\mathrm{RSS} \leftarrow \mathrm{RSS} + \delta_{(k)}^2$
  $\nu \leftarrow k\,d_g - n_\theta$ ▷ positive DOF guaranteed
  $S \leftarrow \log\Gamma\big(\nu/2 + \alpha_0\big) - \big(\frac{\nu}{2} + \alpha_0\big)\log\big(\frac{\mathrm{RSS}}{2} + \beta_0\big) - \frac{\nu}{2}\log(2\pi) - (N{-}k)\,d_g\,\log(2a)$
  if $S > S^*$ **then**
    $S^* \leftarrow S, \quad k^* \leftarrow k$
  end if
end for

4 Experiments

State-of-the-art RANSAC pipelines achieve performance through the combined effect of multiple components: guided sampling (PROSAC [8], NAPSAC [27]), degeneracy testing (DEGENSAC [10], QDEGSAC [15]), local optimization (LO-RANSAC [9, 21]), and varied termination criteria [31]. Many of these components are domain-specific (e.g., guided sampling for two-view matching [3]) and do not transfer to other model classes. Proper ablation of the score from the rest is rarely performed, so a method's apparent advantage may stem from other components rather than the scoring strategy [33]. We follow [33] in isolating the score, with some adjustments called out in the following paragraphs. The takeaway is that every method evaluates the same per-pair pool of minimal-sample hypotheses, differing only in its scoring function and its single hyperparameter.

We evaluate each scoring method along two complementary axes. The *Profile* experiment (Section 4.3) sweeps each method's hyperparameter and measures *robustness to hyperparameter value*, the flatness of the error under perturbation of the optimal parameter on the validation set. The *Sensitivity* experiment (Section 4.4) varies only the validation-set size and measures the smallest validation set sufficient to reach near-optimal test performance.

4.1 Evaluation Protocol

For each image pair we sample $1\,000$ two-view geometric relations (homography, essential, or fundamental matrix, depending on the dataset) from pre-extracted feature correspondences. Each method scores this shared model pool; the highest-scored model is evaluated against the ground-truth transformation by *orientation error*, which is the $\mathrm{SO}(3)$ angular distance between the recovered and oracle rotations. The rotation is recovered per model class: directly from the essential matrix, and via the known calibration for the fundamental matrix ($\mathbf{F} \to \mathbf{E}$) and the homography (by decomposing $\mathbf{H}$). Both Profile and Sensitivity use this same metric. Datasets group image pairs by scene. The Sensitivity experiment (Section 4.4) draws its validation/test split hierarchically by scene on HEB—resampling sites, then pairs within each, to test cross-scene generalization—and by a flat pair-level split on PhotoTourism and ETH3D; the Profile bands (Section 4.3) are pair-level on all three (Table 5).

Shekhovtsov [33] showed that the four threshold-based scoring functions we compare—RANSAC [14], MSAC [35], GaU [37], and MAGSAC++ [5]—can each be reparameterized with a single hyperparameter. Our three proposed marginal-likelihood variants (Jeffreys, IG-Robust, IG-MLE) likewise share a single hyperparameter, the outlier half-width from (8). We write κ for each method's single hyperparameter generically (the half-width a for our profile-marginal scores, the threshold quantile for the baselines), normalized to its per-dataset optimum $\kappa^\star$ in all plots. However, the physical interpretation of the role of the hyperparameter in each family is different, and we should not expect them to share the same numeric range nor do we constrain them to.

Shekhovtsov [33] hand-sets per-method search intervals and shows that even a generous 2+ decade interval based on a reasonable guess can disadvantage some methods, since the interval acts as a weak hyperprior on the hyperparameter. We instead derive every method's interval from data: we draw a single validation set per dataset and evaluate each method's validation-set orientation error on a coarse log-spaced grid; the minimum-error grid point becomes the anchor $\kappa^\star$ (if $\kappa^\star$ lies on a boundary, the search extends by two more decades). The two-decade window $[\kappa^\star/10,\ \kappa^\star \cdot 10]$ is the per-method, per-dataset interval used for hyperparameter selection in both the *Profile* (Section 4.3) and *Sensitivity* (Section 4.4) experiments. To parallelize scoring we sample 100 points per decade uniformly in log-space within this window (200 points total); hyperparameter selection in both experiments is done over this fixed grid. All seven methods use the same resolution and the same two-decade width.

Our empirical-Bayes variants (IG-MLE and IG-Robust) need ground-truth annotations—the relative pose and a set of ground-truth point matches per pair—during the validation phase. Both fit the Inverse-Gamma prior on the noise variance σ^2 from per-pair noise-variance estimates $\hat{\sigma}_p^2$ on the validation set (Section C). *IG-MLE* fits the prior by type-II maximum likelihood: since $\sigma^2 \sim \mathrm{IG}(\alpha, \beta) \iff 1/\sigma^2 \sim \mathrm{Gamma}(\alpha, 1/\beta)$, it fits a Gamma to $\{1/\hat{\sigma}_p^2\}$ and reads off (α_0, β_0), capturing both the location and the dispersion of the $\hat{\sigma}_p^2$ but remaining sensitive to outlier per-pair estimates. *IG-Robust* instead fixes $\alpha_0 = 1$ (two pseudo-observations) and sets $\beta_0 = \alpha_0 \,\mathrm{median}(\hat{\sigma}_p^2)$, trading dispersion information for robustness to the heavy $\hat{\sigma}_p^2$ tails that arise when match-set cardinality approaches the model's degrees of freedom (demonstrated on HEB [6] in Section 4.4). Each $\hat{\sigma}_p^2$ is computed from residuals on that pair's GT matches under the GT pose; how those annotations are obtained is dataset-specific and detailed in Section I.

4.2 Datasets

We evaluate on three datasets, one per two-view geometric model class, with model proposals generated from both a hand-crafted and a state-of-the-art learned feature detector (nearly $70\,000$ pairs in total), spanning three axes of variation: dataset, geometric problem, and feature pipeline. *PhotoTourism* (PT) [19, 24] supplies 21 landmark scenes (44 211 pairs) with RootSIFT [1] features for essential-matrix ($\mathbf{E}$) estimation; *ETH3D* [32] supplies 13 sub-scenes (3 039 pairs) with learned SuperPoint + LightGlue [11, 22] features for fundamental-matrix ($\mathbf{F}$) estimation; and the *Homography Estimation Benchmark* (HEB) [6] supplies 10 architectural sites (21 999 pairs) with SIFT [23] features

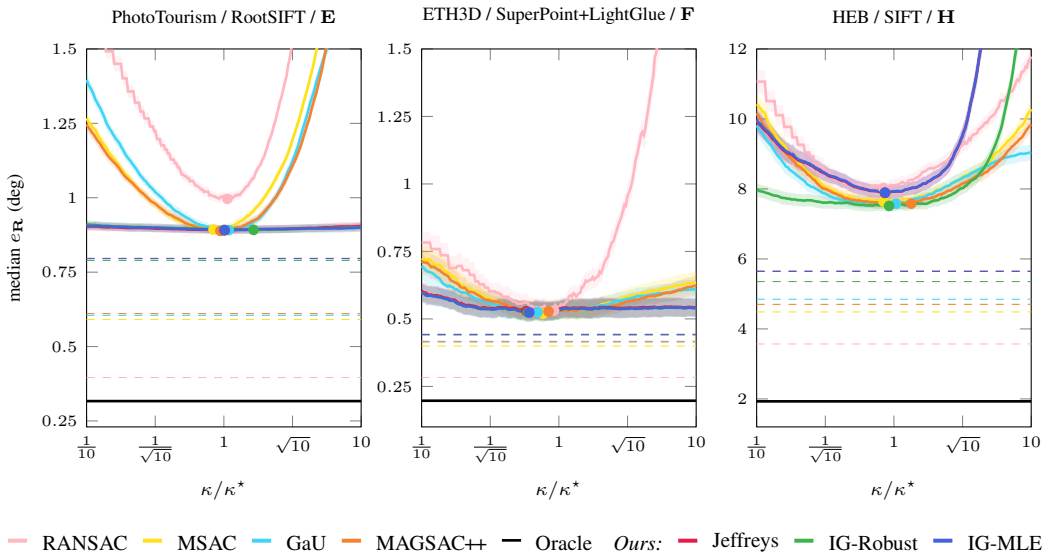


Figure 1: **Profile** across three datasets, one per geometric problem: PT / **E** (left), ETH3D / **F** (middle), HEB / **H** (right). Each panel plots, as solid lines, the median rotation error across all pairs in the dataset versus each method's hyperparameter on a log axis, swept exhaustively over a two-decade interval. Color-matched dashed horizontal lines mark each method's achievable performance when its hyperparameter is free to be tuned per pair to select the model closest to the ground truth; a large gap to the solid curve indicates hyperparameter-dependent model switching. Filled circles mark each method's minimum on the fine grid; the axis is normalized by the coarse-search optimum, so $x = 1$ need not be a minimum. Shaded bands are 95% bootstrap confidence intervals on the median. **In the data-rich regime (PT, ETH3D) our variants are remarkably robust to hyperparameter miscalibration**, with curves nearly flat across the full two-decade window where threshold-based baselines degrade markedly. On HEB—the data-scarce regime—**the IG-Robust prior regularizes the score and outperforms the baselines over most of the hyperparameter interval** (cf. Section 3.2).

for homography (**H**) estimation. The empirical-Bayes variants (IG-MLE, IG-Robust) additionally consume per-pair ground-truth matches during validation (Section 4.1). Full dataset provenance, annotation and feature-pipeline details, and the HEB plane-assignment caveat that surfaces in some HEB panels below are given in Section I (Table 5).

4.3 Profile

The Profile experiment plots each method's median rotation error across all pairs in the dataset as a function of its hyperparameter, swept exhaustively over the two-decade interval described in Section 4.1. It is not a tuning protocol (Section 4.4 gives a practical one): each method's minimum is its achievable lower bound, and the rate at which error climbs off-optimum is a direct readout of sensitivity to miscalibration. No model is selected, so evaluating on the full pool involves no validation–test leakage.

The panels of Figure 1 plot median error against $\kappa/\kappa^\star$, so the same horizontal decade represents comparable parameter neighbourhoods across methods. The bands come from 1 000 pair-resamples of the dataset; because validation and test share the pool, they measure the precision of the fixed-pool median, not a generalization gap (that is the Sensitivity experiment's job). The solid black line is the GT-Oracle, the lowest error in each pair's pool, i.e. the lower bound for any scoring function.

4.3.1 Results

In the data-rich regime (PhotoTourism and ETH3D), our variants are remarkably robust to miscalibration of their hyperparameter a: the median-error curves are nearly flat across the full two-decade window—varying by $<0.02^\circ$ on PhotoTourism—while the threshold-based baselines degrade as

the hyperparameter is perturbed (on PhotoTourism they span $\sim 0.8^\circ$–2.9° over the window, vanilla RANSAC most sharply). Our three variants (Jeffreys, IG-MLE, IG-Robust) overlap throughout the window: IG-Robust fixes $\alpha_0 = 1$ and IG-MLE fits an α_0 that is likewise small relative to the per-pair inlier count ($n_p \approx 140$ on PT), so the data term dominates the posterior (9) and the informative variants approach the Jeffreys limit (Section 3.2). A held-out version of this experiment, with each hyperparameter tuned on validation scenes and then perturbed by up to $128\times$ on PhotoTourism test scenes, shows the same separation: our score retains 0.490 mAA@10° at $128\times$ inflation, against at most 0.134 for the baselines (Section G).

Convergence at the optimum. Our variants and the threshold-based baselines reach the *same* ceiling on every dataset, despite using different inlier distributions: the marginalized model has a Student-t inlier (from the Gaussian + IG hierarchy after integrating σ^2 analytically), while the threshold-based baselines have a Gaussian inlier at a globally pinned σ. Two effects collapse the gap as n_p grows. First, the per-partition posterior on σ^2, $\mathrm{IG}\big(\alpha_0 + \frac{\nu_p}{2},\, \beta_0 + \frac{\mathrm{RSS}_p}{2}\big)$, has shape $\alpha_0 + \nu_p/2 \to \infty$, so the Student-$t$ inlier has $2\alpha_0 + \nu_p$ degrees of freedom and converges to a Gaussian centered at the per-partition scale estimate $\hat{s}_p$. Second, $\hat{s}_p$ has $\sqrt{2/\nu_p}$ relative standard error across pairs, so it concentrates around a population-typical value. An optimally-tuned κ pins σ globally to that same population-typical value, and the model selected per pair coincides with what the marginal score selects. What no fixed σ can reproduce is the flatness under miscalibration above; on HEB the prior no longer recedes and this flatness does not hold, the regime analyzed next.

Low-cardinality regime. HEB stresses a second axis of difficulty: in addition to a low inlier ratio (which controls RANSAC's iteration count), it has a low cardinality of inlying matches per pair, which controls the statistical reliability of any per-partition scale estimate. Mechanically, its annotations are per-plane rather than per-pair, so a typical pair carries hundreds of geometrically valid correspondences but only those on the annotated dominant plane (median $n_p \approx 10$, a $\sim 4\%$ ratio) are GT inliers. The remaining $\sim 96\%$ are *not* mismatches; rather they are correct matches from competing surfaces, but from the perspective of a particular dominant plane they appear as coherently structured outliers, which violate the uniform-outlier assumption shared by the marginal-scoring and threshold-based families. Imperfect assignment of spatially-verified point matches to dominant planes injects further label noise, raising the floor of the per-pair noise-variance estimate $\hat{\sigma}_p^2$ (the variance of inlier residuals under the oracle pose, used to anchor the IG prior). In this context, IG-Robust pulls away from the baselines and the other variants over most of the two-decade search region. With $\alpha_0 = 1$ contributing two pseudo-residuals and a median-anchored β_0, the prior regularizes per-partition scale estimates when n_p is comparable to the model's degrees of freedom. The median anchor is itself insensitive to the inflated $\hat{\sigma}_p^2$ tails produced by structured-outlier contamination and by annotation errors in the dominant-plane labels. MLE regression of the IG parameters is not robust to these corruptions, which negates IG-MLE's regularization advantage. Jeffreys gives up the regularization entirely, and with only $n_p \approx 10$ inliers per pair its per-partition scale estimate is too noisy to discriminate models reliably. Threshold-based methods cannot adapt at all, since a single global σ has no per-partition correction.

Oracle-based scoring. The per-pair-best envelope (dashed line) reports each method's achievable error if its hyperparameter were retuned on every pair to select the model closest to the ground truth. These envelopes invert the ordering of the solid curves' minima: RANSAC gives the best lower bound, threshold-based baselines next, our variants highest. This is not a measure of attainable performance but an artifact of how each family's hyperparameter interacts with the per-pair argmax. For threshold-based methods the hyperparameter is a *model-selection* knob: different thresholds induce different inlier sets and select different models, so a per-pair sweep explores a wide range of selections. For our IG marginal, a enters only the outlier-penalty term $-(N - n_p)\, d_g \log(2a)$; in the data-rich regime the sort-and-sweep picks an inlier partition at the natural residual gap, largely insensitive to a, so sweeping a shifts the score's value but not which model is selected, and the IG per-pair-best collapses to a single fixed model. This decoupling is the desired property: a calibrates score values while the selected model is governed by the data, not the hyperparameter. The threshold-based baselines' lower dashed envelope is therefore not a deployable benefit—it is the visible footprint of the same dual-role mechanism that makes their solid curves so sensitive to hyperparameter perturbation. In deployment one hyperparameter is used per method, so the relevant comparison is between the solid-curve minima; the dashed envelope is a per-pair oracle no real scoring function has access to.

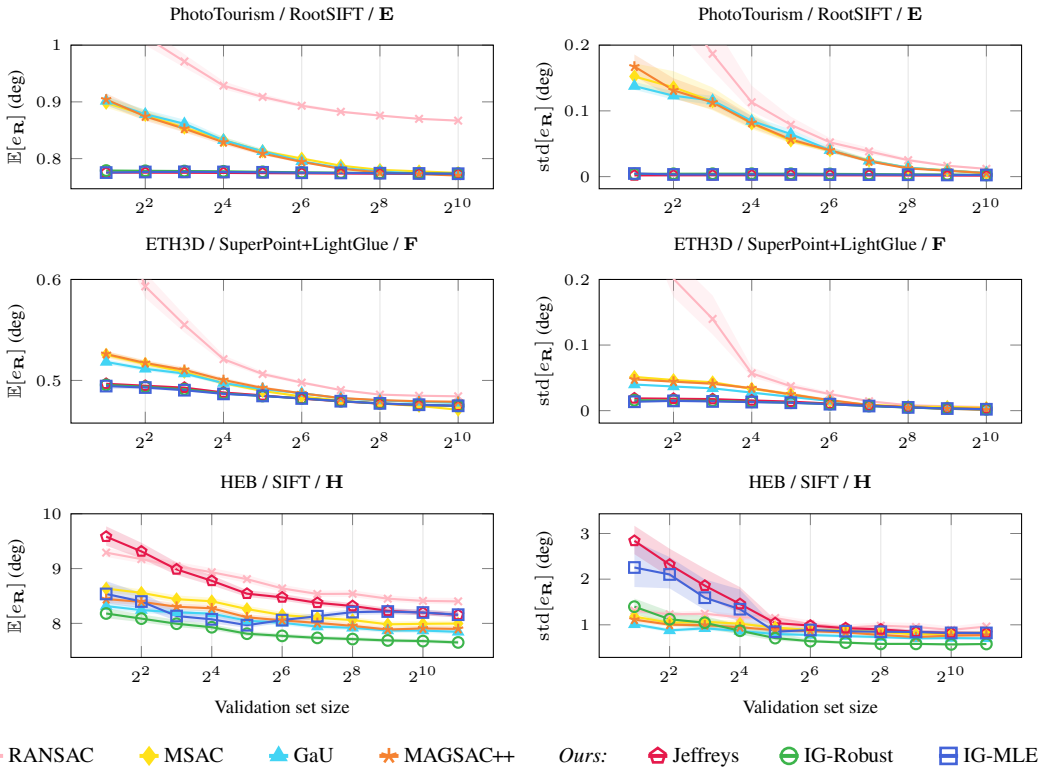


Figure 2: **Sensitivity**: expected test rotation error (left) and its standard deviation (right) as a function of validation-set size, for three cells spanning the geometric-problem axis—PT / **E** (top), ETH3D / **F** (middle), HEB / **H** (bottom). Shaded bands are 95% BCa bootstrap confidence intervals over the cross-validation trials. **Our methods are nearly insensitive to validation-set size**: even with as few as 2–4 validation pairs the IG prior regularizes the per-partition scale enough to recover near-optimal hyperparameters, while threshold-based baselines need 128× more validation data to converge on PhotoTourism (8× on ETH3D at a 5% tolerance). On HEB—where per-pair inlier counts are too sparse to support a stable joint-MLE prior fit—**IG-Robust's median-only anchor** is the appropriate variant.

4.4 Sensitivity

A RANSAC user typically has only a few ground-truth examples available for tuning and selecting their method's hyperparameter. Small validation sets do not fully capture the test distribution, so one scoring method may appear better than another by chance, depending on which validation pairs were drawn [33]. We measure *expected* test performance by randomly drawing small validation sets from the data.

We compute this expected test performance via bootstrap. For each image pair we precompute the rotation error of every method at every grid point in the two-decade hyperparameter window from Section 4.1: 1 000 minimal-sample models are drawn per pair, scored at every hyperparameter κ, the best-scoring model is selected at each κ, and its error against the GT pose is recorded. With these errors cached, the bootstrap loop is fast: draw a validation set $\mathcal{V}$ of cardinality $|\mathcal{V}|$ (on HEB, sampled from a random subset of scenes held out for validation, with the disjoint remaining scenes forming the test pool, so trial variance reflects cross-scene generalization; split at the pair level on PhotoTourism and ETH3D) and pick the validation-optimal hyperparameter $\kappa^\star(\mathcal{V}) = \arg\min_\kappa \text{median}_{p\in\mathcal{V}}\, e_p(\kappa)$, and evaluate $e^{\text{test}}(\kappa^\star(\mathcal{V})) = \text{median}_{p\in\mathcal{T}}\, e_p(\kappa^\star(\mathcal{V}))$ on the held-out test pool $\mathcal{T}$. Repeating 1 000 times gives the expected median test error $\mathbb{E}[e_{\mathbf{R}}]$ at each validation budget $|\mathcal{V}| \in \{2, 4, 8, \ldots, 1024\}$ (extended to 2048 on HEB).

4.4.1 Results

Our variants reach near-optimal expected test error at validation budgets as small as $|\mathcal{V}| = 2$, while the threshold-based baselines need $128\times$ more validation data to converge on PhotoTourism ($8\times$ on ETH3D at a 5% tolerance). On PhotoTourism the three IG variants (Jeffreys, IG-MLE, IG-Robust) sit within 0.006° of their large-$|\mathcal{V}|$ asymptote of 0.77° already at $|\mathcal{V}| = 2$, whereas MSAC, GaU, and MAGSAC++ start at $\sim 0.90^\circ$ and need $|\mathcal{V}| \approx 1024$ to reach the same floor; vanilla RANSAC starts higher (1.08°) and never reaches it, plateauing near 0.87°. ETH3D shows the same convergence pattern. Our variants are also more reproducible at small $|\mathcal{V}|$: their cross-trial standard deviation $\mathrm{std}[e_{\mathbf{R}}]$ sits at 0.002°–0.019° at $|\mathcal{V}| = 2$, while threshold-based baselines disperse over 0.04°–0.38° at the same budget—every baseline at least $2\times$ (ETH3D) to $29\times$ (PhotoTourism) more than any proposed variant.

This is the deployment counterpart of the Profile result: our variants' curves in Figure 1 are nearly flat, so the $\kappa^\star(\mathcal{V})$ chosen on a tiny validation set lands in the wide "good κ" region with high probability, whereas the baselines' steeper curves leave a narrow one that a small $\mathcal{V}$ rarely hits. For the RANSAC practitioner this means our variants can be deployed with a handful of ground-truth pairs—or, in the data-rich regime, with the prior alone—while threshold-based scoring demands a substantial labelled validation set to be tuned reliably.

HEB inverts the prior-irrelevance story. On HEB, Jeffreys is the worst of the three variants at small $|\mathcal{V}|$ (9.58° at $|\mathcal{V}| = 2$ vs. 8.18° for IG-Robust), with the largest trial-to-trial std of any method (2.84°): this is the low-ν_p regime of Section 4.3, where a per-partition scale estimate from $n_p \approx 10$ inliers cannot discriminate models without prior regularization. IG-Robust's median-anchored prior supplies it without being dragged by the heavy $\hat{\sigma}_p^2$ tails that corrupt IG-MLE's joint fit, and dominates HEB at every $|\mathcal{V}|$ in mean (7.65° asymptote vs. 7.85° for the best threshold-based baseline) and in std (lowest of any method for $|\mathcal{V}| \geq 32$).

Complete pipeline. The selected prefix also supplies the inlier ratio and scale that local optimization and termination need (Section H); in one complete RANSAC pipeline (Section G.2) the score is more accurate than SupeRANSAC [2] under every tested scale miscalibration, and faster at calibration (50 versus 111 ms per pair) and under inflation of the scale.

5 Conclusion

We marginalize the inlier scale analytically out of the Gaussian–uniform mixture under a conjugate Inverse-Gamma prior, yielding a closed-form RANSAC score, computable by sort-and-sweep, in which the inlier scale is genuinely absent and a single expression spans the Jeffreys limit and informative empirical-Bayes priors. In the data-rich regime the score is nearly insensitive to hyperparameter miscalibration where threshold-based scoring degrades (Figure 1) and reaches near-optimal test error from two validation pairs (validation selects each method's hyperparameter; the disjoint test split is used only for evaluation) where the baselines need $128\times$ more (Figure 2); in the data-scarce regime the prior supplies regularization that is automatically downweighted as data grows. For the practitioner, a handful of ground-truth pairs, or none where Jeffreys suffices, is enough to deploy near-optimally.

Limitations.
The score is inlier-scale-free, not hyperparameter-free: the outlier half-width a remains, though it is far less sensitive than a threshold (Eq. (11), Section 4.3). The informative variants need annotated correspondences to fit their prior; Jeffreys needs none but is less reliable when the selected inlier set is small. The selected degrees of freedom $\nu^\star$ flag this at inference, since the posterior of σ^2 has relative spread about $\sqrt{2/\nu^\star}$, though a large $\nu^\star$ does not detect structured outliers. The Gaussian–uniform model does not describe structured outliers such as HEB's competing planes. Our evaluation is limited to two-view geometry and mostly to fixed candidate pools.

Acknowledgments and Disclosure of Funding

This work was supported by Kiel Training for Excellence, an EU Horizon Europe programme under the Marie Skłodowska-Curie Actions (MSCA), grant agreement No 101081480, through funding of James Pritts, and by KMS Kiel Marine Science – Centre for Interdisciplinary Marine Science at Kiel University. The authors declare no competing interests.

References

[1] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, pages 2911–2918, 2012.

[2] Daniel Barath. SupeRANSAC: One RANSAC to rule them all. *arXiv:2506.04803*, 2025.

[3] Daniel Barath, Maksym Ivashechkin, and Jiří Matas. Progressive NAPSAC: Sampling from gradually growing neighborhoods. *arXiv:1906.02295*, 2019.

[4] Daniel Barath, Jiří Matas, and Jana Noskova. MAGSAC: Marginalizing sample consensus. In *CVPR*, pages 10189–10197, 2019.

[5] Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiří Matas. MAGSAC++, a fast, reliable and accurate robust estimator. In *CVPR*, pages 1301–1309, 2020.

[6] Daniel Barath, Dmytro Mishkin, Michal Polic, Wolfgang Förstner, and Jiří Matas. A large-scale homography benchmark. In *CVPR*, pages 21360–21370, 2023.

[7] Wojciech Chojnacki, Michael J. Brooks, Anton van den Hengel, and Darren Gawley. On the fitting of surfaces to data with covariances. *IEEE TPAMI*, 22(11):1294–1303, 2000. doi: 10.1109/34.888714.

[8] Ondřej Chum and Jiří Matas. Matching with PROSAC—progressive sample consensus. In *CVPR*, pages 220–226, 2005.

[9] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized RANSAC. In *DAGM*, pages 236–243, 2003.

[10] Ondřej Chum, Tomáš Werner, and Jiří Matas. Two-view geometry estimation unaffected by a dominant plane. In *CVPR*, pages 772–779, 2005.

[11] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *CVPRW*, 2018.

[12] Johan Edstedt. Less biased noise scale estimation for threshold-robust RANSAC. In *CVPRW*, pages 2836–2845, 2025.

[13] Bradley Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.

[14] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[15] Jan-Michael Frahm and Marc Pollefeys. RANSAC for (quasi-)degenerate data (QDEGSAC). In *CVPR*, pages 453–460, 2006.

[16] Matthew Harker and Paul O'Leary. First order geometric distance (the myth of Sampsonus). In *BMVC*, pages 87–96, 2006.

[17] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2003.

[18] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A*, 186(1007):453–461, 1946.

[19] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiří Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *IJCV*, 129(2): 517–547, 2021.

[20] Viktor Larsson and contributors. PoseLib – minimal solvers for camera pose estimation. `https://github.com/PoseLib/PoseLib`, 2020.

[21] Karel Lebeda, Jiří Matas, and Ondřej Chum. Fixing the locally optimized RANSAC. In *BMVC*, pages 95.1–95.11, 2012.

[22] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *ICCV*, pages 17627–17638, 2023.

[23] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[24] Dmytro Mishkin. RANSAC 2020 tutorial starter pack, 2020. `https://github.com/ducha-aiki/ransac-tutorial-2020-data`.

[25] Lionel Moisan and Bérenger Stival. A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. *IJCV*, 57(3):201–218, 2004.

[26] Lionel Moisan, Pierre Moulon, and Pascal Monasse. Automatic homographic registration of a pair of images, with a contrario elimination of outliers. *Image Processing On Line*, 2:56–73, 2012.

[27] D. R. Myatt, P. H. S. Torr, S. J. Nasuto, J. M. Bishop, and R. Craddock. NAPSAC: High noise, high dimensional robust estimation. In *BMVC*, pages 458–467, 2002.

[28] Rahul Raguram and Jan-Michael Frahm. RECON: Scale-adaptive robust estimation via residual consensus. In *ICCV*, pages 1299–1306, 2011.

[29] Felix Rydell, Angélica Torres, and Viktor Larsson. Revisiting Sampson approximations for geometric estimation problems. In *CVPR*, pages 4990–4998, 2024.

[30] Paul D. Sampson. Fitting conic sections to "very scattered" data: An iterative refinement of the Bookstein algorithm. *Computer Graphics and Image Processing*, 18(1):97–108, 1982.

[31] Johannes L. Schönberger, Viktor Larsson, and Marc Pollefeys. Fixing the RANSAC stopping criterion. *arXiv preprint arXiv:2503.07829*, 2025.

[32] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, pages 2538–2547, 2017.

[33] Alexander Shekhovtsov. RANSAC scoring functions: Analysis and reality check. *arXiv preprint arXiv:2512.19850*, 2025.

[34] Charles V. Stewart. MINPRAN: A new robust estimator for computer vision. *IEEE TPAMI*, 17 (10):925–938, 1995.

[35] Philip Torr and Andrew Zisserman. Robust computation and parametrization of multiple view relations. In *ICCV*, pages 727–732, 1998.

[36] Philip H. S. Torr. Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *IJCV*, 50(1):35–61, 2002.

[37] Philip H. S. Torr and Andrew Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Comput. Vis. Image Underst.*, 78(1):138–156, 2000.

[38] Hanzi Wang and David Suter. Robust adaptive-scale parametric model estimation for computer vision. *IEEE TPAMI*, 26(11):1459–1474, 2004.

[39] Hanzi Wang, Daniel Mirota, and Gregory D. Hager. A generalized kernel consensus-based robust estimator. *IEEE TPAMI*, 32(1):178–184, 2010.

Supplementary Material

A Derivation of the Profile-Marginal Score

For a fixed partition $\mathcal{I}$, multiplying the partition likelihood (4) by the Inverse-Gamma prior $\pi(\sigma^2) \propto (\sigma^2)^{-(\alpha_0+1)} \exp(-\beta_0/\sigma^2)$ and integrating gives

$$p(\mathbf{r} \mid \boldsymbol{\theta}, \mathcal{I}) \;\propto\; \frac{1}{(2a)^{n_O d_g}\,(2\pi)^{n_I d_g/2}} \int_0^\infty \frac{\exp\!\big(-\frac{\mathrm{RSS}_\mathcal{I}/2+\beta_0}{\sigma^2}\big)}{(\sigma^2)^{n_I d_g/2+\alpha_0+1}}\, d\sigma^2, \tag{12}$$

where $\mathrm{RSS}_\mathcal{I} \coloneqq \sum_{i\in\mathcal{I}} \|\mathbf{r}_i\|^2$ is the residual sum of squares over inliers and the proportionality absorbs the prior normalizer $\beta_0^{\alpha_0}/\Gamma(\alpha_0)$, which is constant in $(\boldsymbol{\theta}, \mathcal{I})$ and drops out of the partition optimization. The integral evaluates to the Inverse-Gamma normalizing constant

$$\int_0^\infty (\sigma^2)^{-(n_I d_g/2+\alpha_0+1)} \exp\!\Big(-\frac{\mathrm{RSS}_\mathcal{I}/2+\beta_0}{\sigma^2}\Big)\, d\sigma^2 = \Gamma\Big(\frac{n_I\, d_g}{2}+\alpha_0\Big)\Big(\frac{\mathrm{RSS}_\mathcal{I}}{2}+\beta_0\Big)^{-(n_I d_g/2+\alpha_0)}. \tag{13}$$

At $(\alpha_0, \beta_0) = (0, 0)$ this collapses to $\Gamma(n_I d_g/2)\,(\mathrm{RSS}_\mathcal{I}/2)^{-n_I d_g/2}$, the standard Jeffreys-marginalized form.

Replacing $n_I\, d_g$ by the effective degrees of freedom $\nu = n_I\, d_g - n_\theta$ (the n_θ parameters consumed by the minimal-sample fit) and taking logarithms gives the boxed closed form (8); Section A.1 derives this correction.

The variance estimate $\hat{s}^2$. Under the posterior $\sigma^2 \mid \mathcal{I} \sim \mathrm{IG}(\alpha, \beta)$ of (9), with $\alpha = \alpha_0 + \nu/2$ and $\beta = \beta_0 + \mathrm{RSS}_\mathcal{I}/2$, the precision σ^{-2} is Gamma-distributed with shape α and rate β. Equivalently, the same normalizing integral as in (13), with the exponent raised by one, gives $\mathbb{E}[\sigma^{-2} \mid \mathcal{I}] = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+1)}{\beta^{\alpha+1}} = \alpha/\beta$, so $\hat{s}^2 = \beta/\alpha = (2\beta_0 + \mathrm{RSS}_\mathcal{I})/(2\alpha_0 + \nu)$. Unlike the posterior mean $\beta/(\alpha-1)$ and mode $\beta/(\alpha+1)$, it carries no offset and so equals the mean square of the real and pseudo-observations exactly; it is also the variance that reproduces the posterior-averaged Gaussian exponent, $\mathbb{E}[\mathrm{RSS}_\mathcal{I}/2\sigma^2 \mid \mathcal{I}] = \mathrm{RSS}_\mathcal{I}/2\hat{s}^2$.

A.1 The Degrees-of-Freedom Correction

The score (8) uses $\nu = n_I\, d_g - n_\theta$ rather than $n_I\, d_g$, although the partition likelihood (4) is written for a fixed model $\boldsymbol{\theta}$. We now derive this correction and state the assumptions under which (8) is the exact log of a well-defined score.

Setting. A minimal sample $\mathcal{S}$, $|\mathcal{S}| = m$, yields the candidate $\hat{\boldsymbol{\theta}}_\mathcal{S} = f(\mathbf{x}_\mathcal{S})$. We assume an *exact minimal solver*, $m\, d_g = n_\theta$ (true for every model in Table 1), so that

$$\mathbf{r}_i(\hat{\boldsymbol{\theta}}_\mathcal{S}) = \mathbf{0}, \qquad i \in \mathcal{S}, \tag{14}$$

with multi-root solvers handled one root at a time. All residuals below are evaluated at $\hat{\boldsymbol{\theta}}_\mathcal{S}$. A partition splits the observations into inliers $\mathcal{I} \supseteq \mathcal{S}$ and outliers $\mathcal{O}$; $\mathcal{R} = \mathcal{I} \setminus \mathcal{S}$ are the $n_I - m$ non-sample inliers. Inlier residuals follow the Gaussian and outlier residuals the uniform component of (2).

The score is a predictive likelihood. Writing $\mathbf{x}_{\setminus\mathcal{S}}$ for the observations outside the sample, the score of a candidate and partition is

$$\Lambda(\mathcal{S}, \mathcal{I}) = p(\mathbf{x}_{\setminus\mathcal{S}} \mid \mathbf{x}_\mathcal{S}, \mathcal{S}, \mathcal{I}). \tag{15}$$

Because (14) forces the candidate to interpolate its own sample, no non-degenerate density can be assigned to $\mathbf{x}_\mathcal{S}$ under its own candidate, so conditioning on the sample is forced: Λ is an out-of-sample predictive score, with the status of leave-m-out cross-validation, not a marginal evidence.

Plug-in. Replacing the posterior over $\boldsymbol{\theta}$ by a point mass at the solver output, $p(\boldsymbol{\theta} \mid \mathbf{x}_\mathcal{S}, \sigma^2) \approx \delta_{\hat{\boldsymbol{\theta}}_\mathcal{S}}(\boldsymbol{\theta})$, while keeping a prior on the noise scale gives

$$\Lambda(\mathcal{S}, \mathcal{I}) \approx \int_0^\infty p(\mathbf{x}_{\setminus\mathcal{S}} \mid \hat{\boldsymbol{\theta}}_\mathcal{S}, \sigma^2, \mathcal{I})\, \pi(\sigma^2)\, d\sigma^2, \tag{16}$$

which amounts to the working assumption $\mathbf{r}_i(\hat{\boldsymbol{\theta}}_\mathcal{S}) \mid \sigma^2 \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{d_g})$ for $i \in \mathcal{R}$.

Conditioning out the sample. Given $\mathbf{x}_\mathcal{S}$, both $\hat{\boldsymbol{\theta}}_\mathcal{S}$ and, by (14), $\mathbf{r}_\mathcal{S} = \mathbf{0}$ are deterministic; only $\mathbf{r}_\mathcal{R}$ is random. For every bounded g,

$$\mathbb{E}\big[g(\mathbf{r}_\mathcal{S}, \mathbf{r}_\mathcal{R}) \mid \mathbf{x}_\mathcal{S}, \sigma^2, \mathcal{I}\big] = \int g(\mathbf{0}, \mathbf{r}_\mathcal{R})\; p(\mathbf{r}_\mathcal{R} \mid \hat{\boldsymbol{\theta}}_\mathcal{S}, \sigma^2, \mathcal{I})\, d\mathbf{r}_\mathcal{R}. \tag{17}$$

The sample coordinates therefore enter with weight exactly one—not with the density value $(2\pi\sigma^2)^{-m d_g/2}$, since they were forced to zero by the fit rather than observed there—and contribute nothing to the residual sum of squares: $\mathrm{RSS}_\mathcal{I} = \mathrm{RSS}_\mathcal{R}$.

Likelihood. With $n_O = N - n_I$ outliers,

$$p(\mathbf{r}_\mathcal{R}, \mathbf{r}_\mathcal{O} \mid \hat{\boldsymbol{\theta}}_\mathcal{S}, \sigma^2, \mathcal{I}) = (2\pi\sigma^2)^{-\nu/2} \exp\Big(-\frac{\mathrm{RSS}_\mathcal{I}}{2\sigma^2}\Big)\,(2a)^{-n_O d_g}, \quad \nu = (n_I - m)\, d_g = n_I\, d_g - n_\theta, \tag{18}$$

using $m\, d_g = n_\theta$: exactly the n_θ Gaussian coordinates consumed by the fit are removed, while the RSS is unchanged.

Marginalization. With $\sigma^2 \sim \mathrm{IG}(\alpha_0, \beta_0)$, $\alpha_0, \beta_0 > 0$, the Inverse-Gamma integral (13) at $\alpha = \alpha_0 + \nu/2$, $\beta = \beta_0 + \mathrm{RSS}_\mathcal{I}/2$ gives

$$\Lambda(\mathcal{S}, \mathcal{I}) = (2\pi)^{-\nu/2}\,(2a)^{-n_O d_g}\,\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\,\frac{\Gamma(\alpha_0 + \nu/2)}{(\beta_0 + \mathrm{RSS}_\mathcal{I}/2)^{\alpha_0 + \nu/2}}, \tag{19}$$

whose logarithm is (8) up to the prior normalizer $\beta_0^{\alpha_0}/\Gamma(\alpha_0)$, which is common to all partitions and candidates. The score is finite for every partition: $\nu = 0$ gives the all-outlier baseline $(2a)^{-(N-m)d_g}$, and $\mathrm{RSS}_\mathcal{I} \to 0$ stays bounded because $\beta_0 > 0$. Letting $\alpha_0, \beta_0 \to 0$ and dropping the improper constant, again common to all partitions so that the ranking is preserved, recovers the Jeffreys form (10),

$$\Lambda_J(\mathcal{S}, \mathcal{I}) \;\propto\; \frac{\Gamma(\nu/2)}{(\pi\ \mathrm{RSS}_\mathcal{I})^{\nu/2}}\,(2a)^{-n_O d_g}, \qquad \nu > 0,\ \mathrm{RSS}_\mathcal{I} > 0. \tag{20}$$

Scope. The argument requires exact minimal candidates. When d_g does not divide n_θ (e.g. a homography with one distortion parameter, $n_\theta = 9$, $d_g = 2$), the minimal sample over-determines the model, its residuals are not all zero, and the interpolation step (14) fails. After local optimization (14) no longer holds, $\mathrm{RSS}_\mathcal{I} = \mathrm{RSS}_\mathcal{R}$ fails with it, and the same count $\nu = n_I\, d_g - n_\theta$ follows instead from the classical projection argument for a least-squares fit on $\mathcal{I}$.

B The Score's Implicit Threshold

We derive the implicit threshold (11) for general codimension d_g and prior (α_0, β_0). Write $\nu_k = k\, d_g - n_\theta$, $\mathrm{RSS}_k = \sum_{j=1}^{k} \delta_{(j)}^2$, $A_k = \alpha_0 + \nu_k/2$ and $B_k = \beta_0 + \mathrm{RSS}_k/2$ for the prefix $\mathcal{I}_k$ of Algorithm 1, so that $\hat{s}_k^2 = B_k/A_k$ is the variance estimate of Section 3.2.

Exact increment. Moving the next point, at distance $\delta = \delta_{(k+1)}$, from the outliers to the inliers adds d_g residual coordinates, adds δ^2 to the residual sum of squares and removes one outlier, so by (8)

$$\begin{aligned}\Delta S_k \coloneqq S(\boldsymbol{\theta}, \mathcal{I}_{k+1}) - S(\boldsymbol{\theta}, \mathcal{I}_k) = \log\frac{\Gamma(A_k + d_g/2)}{\Gamma(A_k)} - \Big(A_k + \frac{d_g}{2}\Big)\log\Big(B_k + \frac{\delta^2}{2}\Big) \\ + A_k \log B_k - \frac{d_g}{2}\log(2\pi) + d_g \log(2a).\end{aligned} \tag{21}$$

Since $-(A_k + \frac{d_g}{2})\log(B_k + \frac{\delta^2}{2}) + A_k \log B_k = -\frac{d_g}{2}\log B_k - (A_k + \frac{d_g}{2})\log(1 + \frac{\delta^2}{2B_k})$, the increment is exactly $\Delta S_k = \log\big(p_{\mathrm{in}}^*(\mathbf{r})/p_{\mathrm{out}}\big)$ at $\|\mathbf{r}\| = \delta$, where

$$p_{\mathrm{in}}^*(\mathbf{r}) = \frac{\Gamma(A_k + d_g/2)}{\Gamma(A_k)\,(2\pi B_k)^{d_g/2}}\Big(1 + \frac{\|\mathbf{r}\|^2}{2B_k}\Big)^{-(A_k + d_g/2)}$$

is the posterior predictive density of a new residual, a d_g-variate Student-t with $2A_k$ degrees of freedom and scale $\hat{s}_k$. This holds for every prefix, including the small ones just beyond the minimal sample.

Large-sample limit. As A_k grows the Student-t tends to the Gaussian: with $\log\Gamma(A_k + d_g/2) - \log\Gamma(A_k) \approx (d_g/2)\log A_k$ and $(A_k + d_g/2)\log(1 + \delta^2/2B_k) \approx A_k\,\delta^2/2B_k$,

$$\Delta S_k \;\approx\; -\frac{d_g}{2}\log\!\big(2\pi\hat{s}_k^2\big) - \frac{\delta^2}{2\hat{s}_k^2} + d_g\log(2a) \;=\; \log\frac{p_{\mathrm{in}}(\mathbf{r};\,\hat{s}_k)}{p_{\mathrm{out}}}\Big|_{\|\mathbf{r}\|=\delta}.$$

Adding a point thus raises the score exactly when the Gaussian with the current variance estimate explains it better than the uniform outlier density, the MAP rule of the Gaussian–uniform mixture at $\gamma = \frac{1}{2}$. The increment decreases in δ, so along the sorted sweep it changes sign once: the optimal prefix ends where $p_{\mathrm{in}}(\mathbf{r};\hat{s}) = p_{\mathrm{out}}$ at $\|\mathbf{r}\| = \tau^*$, with $\delta_{(k^*)} \le \tau^* < \delta_{(k^*+1)}$. Solving for τ^* gives (11). For finite ν the Student-t's heavier tails place the exact boundary slightly further out.

C Estimating the Prior Parameters (α_0, β_0)

The empirical-Bayes variants fit the prior from per-pair scale estimates $\{\hat{\sigma}_p^2\}_p$ on the validation set. For pair p with labelled inlier set $\mathcal{I}_p^{\mathrm{gt}}$ of size n_p, we fit the model to $\mathcal{I}_p^{\mathrm{gt}}$ by an unconstrained least-squares fit (DLT for $\mathbf{H}$) and set $\hat{\sigma}_p^2 = \mathrm{RSS}_p^{\mathrm{gt}}/(d_g\,n_p - n_\theta)$, with $\mathrm{RSS}_p^{\mathrm{gt}}$ the Sampson sum of squares of the inliers under that fit. We do not fit under the oracle pose (for $\mathbf{H}$, only the 3-DOF plane normal): HEB's labelled inliers are not strictly coplanar, so that estimate inflates by orders of magnitude on a non-trivial fraction of pairs. For relative pose the two fits coincide. Pairs with $d_g\,n_p \le n_\theta$ yield no $\hat{\sigma}_p^2$ and are excluded from the prior aggregation; they remain test pairs, since the score (8) only requires the selected partition to satisfy $\nu > 0$.

IG-MLE: type-II maximum likelihood. Using the identity $\sigma^2 \sim \mathrm{IG}(\alpha,\beta) \iff 1/\sigma^2 \sim \mathrm{Gamma}(\alpha,\,1/\beta)$, we fit a Gamma distribution to $\{1/\hat{\sigma}_p^2\}_p$ and read off (α_0,β_0) from its shape and inverse scale. Both parameters are jointly identified, so the prior captures both the location and the dispersion of the validation $\hat{\sigma}_p^2$. This estimator is preferred when the per-pair $\hat{\sigma}_p^2$ are well concentrated, but is sensitive to outliers: a pair with few inliers gives a $\hat{\sigma}_p^2$ with relative standard error $\sqrt{2/\nu_p^{\mathrm{gt}}}$, $\nu_p^{\mathrm{gt}} = d_g\,n_p - n_\theta$, and a small number of extreme values can drag the joint MLE far from the bulk of the distribution.

IG-Robust: median anchor. We fix $\alpha_0 = 1$, i.e. two pseudo-observations, and set $\beta_0 = \alpha_0 \cdot \mathrm{median}(\hat{\sigma}_p^2)$, so that the prior mean square $s_0^2 = \beta_0/\alpha_0$ is the median of $\{\hat{\sigma}_p^2\}$. This sacrifices the dispersion information captured by joint MLE in exchange for robustness to heavy tails in $\{\hat{\sigma}_p^2\}$, which arise on datasets with widely-varying per-pair inlier counts.

Both estimators are recomputed once per cross-validation trial and held fixed for both the a-selection sweep and the test evaluation.

D Relation to Torr's Bayesian Framework, GRIC, and MAGSAC++

Torr's Bayesian manifold formulation. Torr's Bayesian treatment of epipolar geometry and generic manifold fitting [36] already shows that marginalizing the latent noiseless correspondence removes the tangential error, leaving a Gaussian likelihood in the manifold codimension d_g, together with a uniform outlier model. This is the foundation our likelihood (2) builds on.

GRIC. GRIC addresses *model-class* selection: it approximates the evidence for manifolds of different dimensions and parameter counts using a fixed-scale MAPSAC/MSAC data term plus complexity penalties for the latent point coordinates and the global model parameters. In our setting the model class is fixed, so these penalties are constant across hypotheses, while the data term still requires an externally specified σ. Our contribution is complementary: starting from the same codimension-based Gaussian–uniform likelihood, we condition on the candidate and its minimal sample, profile over the inlier partition, and integrate σ^2 exactly under an Inverse-Gamma prior. GRIC approximates model-class evidence; our score removes the inlier scale for within-class RANSAC scoring. Our $\nu = n_I\,d_g - n_\theta$ counts residual Gaussian evidence after fitting (Section A.1); it is not GRIC's complexity penalty.

The MAGSAC++ likelihood. MAGSAC++ [5] begins from a χ_4 density because a two-view correspondence has four coordinates. The residual, however, is a distance to the model manifold, so only the d_g orthogonal constraint directions remain ($d_g = 1$ for essential and fundamental matrices, $d_g = 2$ for homographies): the relevant dimension is the model codimension, not the ambient observation dimension (Section E). More fundamentally, even χ_{d_g} is not an appropriate model-ranking likelihood. It is the density of the *radius* of a Gaussian residual vector and includes the volume factor r^{d_g-1}; because $r = r(\boldsymbol{\theta})$ depends on the candidate, this is a parameter-dependent transformation of the observations, and for $d_g \geq 2$ it assigns zero density to an exact fit. Our formulation instead starts from the Gaussian density of the d_g-dimensional residual vector itself: d_g is the per-observation constraint dimension and ν the total residual evidence after fitting, and neither is used as the shape of a radial likelihood.

E Model Dimensions: Parameters, Codimension, and Minimal Sample Size

Table 1 lists the quantities of Section 2: n_θ is the number of free parameters after fixing gauge freedoms, d_g the codimension, $m = \lceil n_\theta / d_g \rceil$ the minimal sample size (the Jacobian $\partial_{\boldsymbol{\theta}} \mathbf{g}_{\mathcal{I}} \in \mathbb{R}^{k\, d_g \times n_\theta}$ generically has rank n_θ only for $k \geq m$), and d the dimension of an observation $\mathbf{x}_i$.

Table 1: Dimensional quantities for standard two-view models. n_θ: model parameters (free DOF); d_g: codimension (constraints per point); $m = \lceil n_\theta / d_g \rceil$: minimal sample size; d: observation dimension.

Model	n_θ	d_g	m	d
Fundamental matrix ($\mathbf{F}$)	7	1	7	4
Homography ($\mathbf{H}$)	8	2	4	4
Essential matrix ($\mathbf{E}$)	5	1	5	4

F Scoring Runtime

Table 2 reports single-threaded per-hypothesis timings on PhotoTourism. All scorers receive identical precomputed residuals. For MAGSAC++ we use the lookup-table implementation prescribed by the paper and shipped with the reference code, its most optimistic faithful runtime; the closed form is included for completeness. Although comparison sorting is $O(N \log N)$, our implementation sorts with a radix sort on 32-bit float keys, so the sort-and-sweep is $\Theta(N)$: increasing the correspondence count by $2.21\times$ increases the runtime by $2.24\times$, against the $2.48\times$ that $N \log N$ scaling would predict. Our score costs 8.8–19.7 μs more per hypothesis than MSAC and 7.8–16.6 μs more than the MAGSAC++ lookup table.

Table 2: Per-hypothesis scoring time (single thread, identical precomputed residuals) at two mean correspondence counts $\bar{n}$.

Method	Form	$\bar{n} = 715$	$\bar{n} = 1580$
RANSAC	exact	0.65 μs	1.43 μs
MSAC	exact	0.84 μs	1.88 μs
MAGSAC++	lookup table	1.83 μs	5.00 μs
Ours	exact sort-and-sweep	9.62 μs	21.59 μs
GaU	exact	12.38 μs	28.49 μs
MAGSAC++	closed form	29.73 μs	92.99 μs

These microbenchmarks are not end-to-end slowdowns. A complete RANSAC iteration also includes minimal solving, residual evaluation and local optimization, and the termination criterion determines how many times all of these costs are paid. Section G.2 therefore measures a complete pipeline.

G Scale Miscalibration

The main experiments isolate the scoring function on fixed candidate pools with each method's hyperparameter tuned on the data. This section measures how the scores (Section G.1), and a complete RANSAC pipeline built on ours (Section G.2), degrade when the user-supplied scale is wrong.

Scale corruption. Methods are tuned once at calibration. At test time, each user-supplied quantity expressed in residual units is multiplied by $\zeta \in [1/128, 128]$ (deflation for $\zeta < 1$, inflation for $\zeta > 1$); dimensionless multipliers and scales recomputed from the selected partition are unchanged. Thus ζ measures sensitivity to external calibration rather than perturbing every internal variable identically.

Metric. Unlike the main experiments, which report the orientation error (Section 4.1), this section reports the pose error, which also includes the translation direction. For relative pose, we evaluate the four essential-matrix decompositions and report the smallest maximum angular error [19]. We define mAA@10° as

$$\begin{aligned} \mathrm{mAA@10^\circ} &= \frac{1}{10^\circ}\int_0^{10^\circ} P(e_{\mathrm{pose}} \le t)\, dt, \\ \text{s.t.} \quad e_{\mathrm{pose}} &= \min_{j\in\{1,2,3,4\}} \max\big(e_R^{(j)}, e_t^{(j)}\big). \end{aligned} \tag{22}$$

Here $e_R^{(j)}$ and $e_t^{(j)}$ are the rotation and translation-direction errors of decomposition j. In (22), t is the pose-error threshold, and $P(e_{\mathrm{pose}} \le t)$ is the per-scene fraction of pairs whose pose error is at most t. We compute mAA per scene and then average over scenes, with failed estimates contributing zero.

Parameter mapping. The shared scoring coordinate $q \in [1, 3.3]$ and the validation reference scale $\hat{\sigma}$, the square root of the median per-pair noise variance $\hat{\sigma}_p^2$ (Section C) over the validation pairs, determine each score's native residual-unit parameter; q plays the role of the quantile in $\tau = q\sigma$ (Section 1). RANSAC, MSAC, SIMFIT++, and GaU use a threshold or cutoff $\tau = q\hat{\sigma}$. MAGSAC++ uses the upper noise-scale bound $\bar{\sigma} = q\hat{\sigma}$, not a residual cutoff. For the profile-marginal score, the outlier half-width is

$$a(q; \hat{\sigma}, d_g) = \hat{\sigma}\sqrt{\frac{\pi}{2}}\exp\left(\frac{q^2}{2d_g}\right), \tag{23}$$

which equates the isotropic Gaussian inlier density and uniform outlier density at residual norm $q\hat{\sigma}$ under equal assignment priors, the inverse of the implicit threshold (11). This is a calibration map, not the exact boundary of the profile-marginal partition; generatively, a remains an outlier-domain parameter.

Replacing $\hat{\sigma}$ by $\zeta\hat{\sigma}$ at fixed q^* multiplies each native parameter by ζ:

$$\begin{aligned} \tau(q^*; \zeta\hat{\sigma}) &= \zeta\tau(q^*; \hat{\sigma}), \\ \bar{\sigma}(q^*; \zeta\hat{\sigma}) &= \zeta\bar{\sigma}(q^*; \hat{\sigma}), \\ a(q^*; \zeta\hat{\sigma}, d_g) &= \zeta a(q^*; \hat{\sigma}, d_g). \end{aligned} \tag{24}$$

The scoring-only experiment (Section G.1) applies the nearest evaluation-grid value to every test pair without retuning. Inverting the maps at the original reference scale $\hat{\sigma}$ gives the equivalent coordinates

$$q'_\tau = q'_{\bar{\sigma}} = \zeta q^*, \qquad (q'_a)^2 = (q^*)^2 + 2d_g \log\zeta, \tag{25}$$

where q'_a is real only when the right-hand side is nonnegative.

G.1 Scoring

This experiment evaluates how well each score selects a model under scale miscalibration, without local optimization: each score ranks the identical pool of minimal-sample hypotheses of each image pair, and the selected hypothesis is evaluated directly. Both experiments of this section use a separate set of PhotoTourism pools: the 21 scenes are split by scene, not by pair as in Section 4, into 10 validation and 11 test scenes of 2 200 pairs each, with RootSIFT correspondences capped at 4 096 and 4 000 minimal-sample hypotheses per pair. Ours is the profile-marginal score with the Jeffreys prior. Cross-validation uses only the shared scoring window $q \in [1, 3.3]$, a dimensionless multiple

of the validation reference scale $\hat{\sigma}$ (Section G). At test time the supplied scale is multiplied by $\zeta \in [1/128, 128]$ with the tuned coordinate q^* held fixed; each score's parameter follows through its map (Section G) and is used for all pairs without retuning. For the profile-marginal score, the affected parameter is the outlier half-width a; the selected partition and its variance estimate $\hat{s}^2$ remain data-dependent. SIMFIT++ [12] is MSAC with a data-driven threshold: starting from an initial threshold, it alternates fitting on a random half of the correspondences with a bias-corrected median scale estimate on the other half and scores at the converged threshold; its initial threshold is cross-validated on the τ axis like the other threshold-based scores, and the estimate falls back to it when the iteration leaves the admissible range. Because that range is the reference's 0.25–8 px window, SIMFIT++ scores the pool's residuals re-expressed in pixels through the stored image normalizations; every other score works in the normalized frame, where all of its parameters are defined. The shaded confidence intervals in Fig. 3 use a scene-clustered BCa bootstrap [13].

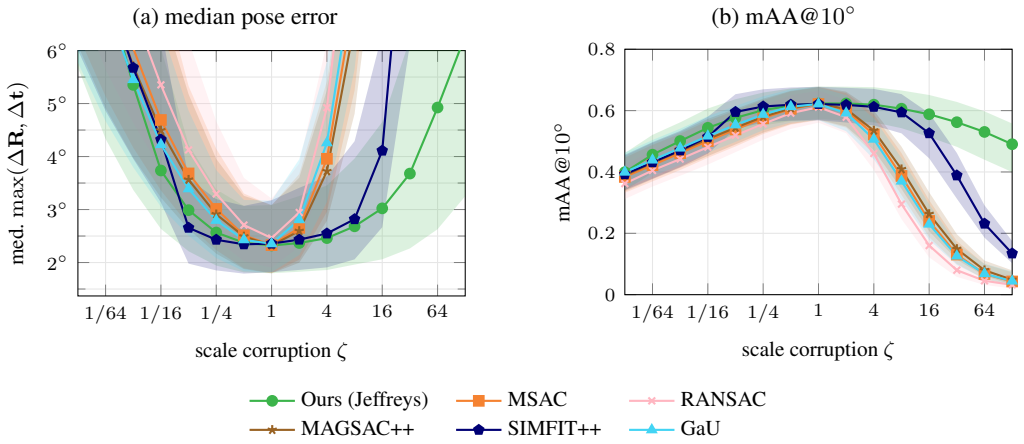


Figure 3: **Scoring under scale miscalibration, without local optimization** (PhotoTourism essential matrices; 11 test and 10 validation scenes). We report (a) the scene-averaged median pose error, with the axis clipped at $6°$ so that the near-tied calibrated regime stays readable (the threshold-based scores leave the panel under inflation and reach $37°$–$45°$ at $128\times$), and (b) mAA@$10°$ (22) at the same operating points, which are tuned on the median. The baselines are RANSAC [14], MSAC [35], MAGSAC++ [5], GaU [33], and SIMFIT++ [12], which re-estimates the inlier threshold per pair from held-out residuals of the hypothesis selected on a random half of the correspondences.

At calibration ($\zeta = 1$) the scores are nearly tied (Fig. 3): the profile-marginal score and SIMFIT++ reach median errors of $2.32°$ and $2.36°$ and mAA of 0.623 and 0.622. At $128\times$ inflation, the profile-marginal score reaches a median error of $6.3°$ and 0.490 mAA, while SIMFIT++ reaches $34°$ and 0.134 mAA; the other threshold-based scores have median errors of $37°$–$45°$ and at most 0.049 mAA. The profile-marginal score is thus substantially more stable under scale corruption. This reflects both its data-dependent partition selection and scale estimation; by (25), the equivalent coordinate of the half-width also changes only with $\log \zeta$, $(q'_a)^2 = (q^*)^2 + 2d_g \log \zeta$, and the experiment does not isolate these effects. Re-estimating the threshold per pair (SIMFIT++) does not remove the dependence on the supplied scale: at $128\times$ inflation its median error of $34°$ is close to that of the other threshold-based scores, against $6.3°$ for the profile-marginal score.

These pools hold more hypotheses than the scores need: the budget of 1000 minimal samples per pair used in Section 4 suffices for every scorer: on 8 800 PhotoTourism pairs with about 4 000 RootSIFT correspondences each, from four held-out scenes, all scorers tie at correct calibration (mAA@$10°$ 0.534–0.542, median error $2.80°$–$2.96°$), and under $32\times$ miscalibration ours retains 0.466 mAA and $4.60°$ while the baselines fall to 0.10–0.14 mAA and $20°$–$25°$.

G.2 End-to-End RANSAC

The deployed pipeline (Ours) combines the profile-marginal score with local optimization whose loss width is anchored to the selected partition (Section H). We compare it with SupeRANSAC [2], including sampling and adaptive termination, on the 11 PhotoTourism test scenes (Table 3). Both are

tuned at calibration; at test time, ζ multiplies each system's residual-unit parameter. At calibration, ours reduces the median pose error from 2.23° to 1.57° and raises mAA@10° from 0.64 to 0.69. Under corruption, our median error stays within 1.8–6.2°, whereas SupeRANSAC's reaches 54° under deflation and 38.27° at $128\times$ inflation. We postulate that the calibrated gain has the same origin: a threshold tuned on validation pairs is effectively miscalibrated for test pairs whose noise scale lies outside the validation distribution, whereas partition anchoring estimates the scale per pair.

Table 3: **RANSAC under scale miscalibration.** Scene-averaged median pose error and mAA@10° on the 11 PhotoTourism test scenes; median wall time per pair, single-threaded, on a stratified sample of 44 pairs per scene (median of 5 repeats); both systems tuned at $\zeta = 1$, SupeRANSAC at its validated maximum noise level with at most 10 000 trials. Bold marks the better system per column.

	ζ	$1/16$	1	4	16	128
Ours	error [$^\circ$]	**2.69**	**1.57**	**1.84**	**2.59**	**6.20**
	mAA@10°	**0.595**	**0.686**	**0.667**	**0.620**	**0.503**
	time [ms]	291	**50**	**34**	**23**	**15**
SupeRANSAC [2]	error [$^\circ$]	53.73	2.23	3.49	10.14	38.27
	mAA@10°	0.210	0.643	0.558	0.333	0.072
	time [ms]	**86**	111	168	265	381

H Local Optimization and Termination

The sort-and-sweep supplies every quantity the downstream components of RANSAC need: the selected support $\mathcal{I}^\star$, its cutoff $\tau^\star = \delta_{(k^\star)}$, the inlier ratio $\hat{w} = k^\star/N$, and the variance estimate $\hat{s}^2$ of Section 3.2. Standard confidence-based termination stops after

$$K = \frac{\log(1-p)}{\log(1-\hat{w}^{\,m})} \tag{26}$$

minimal samples, for a confidence p that an all-inlier sample has been drawn, and local optimization uses either $\tau^\star$ or a partition-anchored width $\lambda\hat{s}$, with only the dimensionless λ cross-validated. No calibrated residual threshold is required at test time.

Table 4 isolates this effect with the same PoseLib truncated-least-squares Levenberg–Marquardt optimizer (TLS-LM) [20], on the PhotoTourism pools of Section G. The fixed variant uses a calibrated threshold τ_{ref}, perturbed at test time by ζ; the anchored variant re-estimates $\tau = \lambda\hat{s}$ for each hypothesis. Both are tuned at $\zeta = 1$.

Table 4: Local optimization with a fixed versus a partition-anchored width. Median pose error [deg] over the 11 PhotoTourism test scenes; ζ multiplies the calibrated residual unit at test time.

PoseLib TLS-LM	$\zeta = 1$	$\zeta = 16$	$\zeta = 128$
fixed (τ_{ref})	1.8	4.0	22.0
anchored ($\tau = \lambda\hat{s}$)	1.9	2.4	4.2

I Dataset Details

Table 5 summarizes the three datasets, which span three axes of variation—dataset, geometric problem, and feature pipeline.

Annotations. Each dataset provides image pairs with ground-truth absolute pose and camera calibration, from which we construct the ground-truth relative pose. HEB additionally annotates the homographies induced by the dominant planes in each scene. PhotoTourism has no canonical split; we use the publicly available scenes with ground-truth poses and SfM matches.

Each scoring method ranks a set of *tentative correspondences* per pair; IG-MLE and IG-Robust additionally consume *ground-truth point matches* during validation (Section 4.1). HEB and PhotoTourism

Table 5: Datasets used in our evaluation: geometric problem, feature pipeline, bootstrap protocol, number of scenes, and total image-pair count.

Dataset	Problem	Features	Protocol	Scenes	Pairs
PhotoTourism [19]	**E**	RootSIFT [1]	Flat (resample)	21	44 211
ETH3D [32]	**F**	SP+LG [11, 22]	Flat (resample)	13	3 039
HEB [6]	**H**	SIFT [23]	Hierarchical	10	21 999

ship both a tentative correspondence set—SIFT [23] on HEB, RootSIFT [1] on PhotoTourism—and SfM bundle-adjusted GT matches (for HEB, restricted to the annotated plane). These GT matches are pseudo-ground-truth, a multi-view SfM reference treated as gold-standard for the two-view subproblem; they also give the per-pair variance estimates $\hat{\sigma}_p^2$ of Section C. ETH3D ships only per-view laser depth (training subset only; the test set is unreleased); we run SuperPoint [11] + LightGlue [22] (SP+LG), a learned detect-and-match pipeline, as the tentative set, and verify each candidate by 3D consistency: pixels are unprojected via the depth, brought into a common frame by the GT poses, and accepted when the resulting 3D points lie within a dataset-specific threshold (reflecting feature, depth-sensor, and absolute-pose accuracy). Pairing SP+LG with SIFT/RootSIFT on the other two datasets exercises our variants against feature noise from both hand-designed and learned pipelines. On HEB a pair's matches can span several planes, so a scoring method may select a model that fits one plane while the GT inliers belong to another; this surfaces in some HEB panels of the experiments.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are derived in Sections 2 and 3 (the profile-marginal score, its closed form, its implicit threshold and its cost) and validated on nearly 70 000 image pairs in Section 4; the scope (inlier-scale-free, not hyperparameter-free) is stated in the introduction and the Limitations paragraph of Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Limitations paragraph of Section 5 states the remaining hyperparameter a, the need for annotated data by the informative variants, the reduced reliability of Jeffreys for small inlier sets (flagged by $\nu^\star$), structured outliers, and the restriction to two-view geometry and mostly fixed candidate pools; Section 4.4 and Section I detail the data-scarce HEB regime and its ground-truth caveat.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Sections 2 and 3 state the assumptions (the Gaussian–uniform mixture, isotropic Gaussian residuals to first order, a conjugate Inverse-Gamma prior, and a plug-in candidate from an exact minimal solver); Section A derives the closed form, Section A.1 the degrees-of-freedom correction and its scope, and Section B the implicit threshold.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 gives the evaluation protocol, Section C the prior estimators, Section I the datasets, and the reproduction project with the configurations of record is available at `https://ransacscoringdoneright.github.io`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The method is evaluated on publicly available data. The code (an open-source Julia package) and a reproduction project with the configurations of record and instructions are available at `https://ransacscoringdoneright.github.io`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 gives the hyperparameter grids, validation and test splits and hypothesis budgets; Sections C, F and I give the remaining details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figure 1 shows 95% bootstrap confidence intervals on the median error, and Figure 2 reports the mean and standard deviation of the test error over bootstrap draws of the validation set.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Sections F and G.2 report single-threaded per-hypothesis and end-to-end timings, but not the hardware or the total compute of the study; the score itself is cheap, and the experiments need no special resources, though they parallelize well.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The presented work conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: We do not expect societal impact of the performed work.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We do not see a risk for misuse of the method.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Utilized datasets and benchmarks are cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code is an MIT-licensed Julia package with documentation at `https://ransacscoringdoneright.github.io`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.