DUSt3R：点图、尺度与多视图几何接口

12 页原创中英双语数学推导与工程映射深读手册

01 / 统一三维接口：先点图，后相机

A01 · 原文 PDF p.1,2,3,4 · 固定版本溯源；本页含原创分析

DUSt3R 把一对 RGB 图像直接回归为两幅逐像素三维点图，并把二者都表达在第一张图的相机坐标系。传统流程通常先匹配像素、估计相机，再由三角化获得三维；这里先预测共享三维几何，再从中提取对应、深度和相机参数。

点图不是无序点云：每个像素保存一个三维坐标，图像索引与几何索引一一对应。它同时保留二维观测位置与三维场景信息，因此可以把匹配变成共享坐标系里的近邻，把深度变成特定相机系的 z，把相机估计变成几何拟合。

底本为 arXiv:2312.14132v3，2024-12-02，23 页，发表于 CVPR 2024。作者为 Shuzhe Wang、Vincent Leroy、Yohann Cabon、Boris Chidlovskii、Jerome Revaud。该 arXiv 版本采用非独占分发许可；本册采用原创双语推导、算例和工程映射，随附完整英文原文。

“无预先标定输入”不等于“输出自动为米制”，也不等于“无需几何验证”。网络学习训练数据中的先验，输出有未知尺度，而且通用架构未硬性强制点图符合针孔相机。对机器人而言，它是强几何初始化与多任务接口，不是默认校准过的安全深度传感器。

$$F(I_1,I_2) = (X^{1,1}, X^{2,1}, C^{1,1}, C^{2,1})$$

$X \in \mathbb{R}^{HxWx3}$；$C \in \mathbb{R}^{HxW}$；output coordinates: camera 1

接口变化	传统常见路线	DUSt3R 路线
先求什么	匹配/相机	共享坐标点图
深度	三角化结果	自身相机系点图的 z
尺度	依赖基线/标定	仍需外部尺度锚

DUSt3R reverses a familiar dependency: it predicts two pixel-aligned pointmaps in the first camera's frame, then derives correspondence, depth and camera parameters. A pointmap retains pixel identity, unlike an unordered point cloud.

Input calibration is not required, but the output has unknown scale and is not explicitly constrained to a physically valid pinhole model. Learned priors make this interface useful without establishing metric accuracy for every scene.

This original companion uses the fixed 23-page arXiv v3 by Wang, Leroy, Cabon, Chidlovskii and Revaud. The non-exclusive license selects an analytical handbook rather than complete translated reproduction.

02 / 点图坐标：第二幅的 z 不是第二相机深度

A02 · 原文 PDF p.3,4,16 · 固定版本溯源；本页含原创分析

对于针孔相机，像素 (u,v) 的轴向深度为 Z 时，三维点为 $ZK^{-1}(u,v,1)^{T}$。这里 Z 是相机 z 坐标，不是相机中心到三维点的欧氏距离。偏离主轴时，range=Z||$K^{-1}q$||，若把二者混用，会产生随视角增加的系统性误差。

上标 Xn,m 表示第 n 张图的像素所对应的三维点，用第 m 个相机坐标系表示。若 Tcw 是世界到相机的齐次变换，则 $Xn,m=Tm,w(Tn,w)^{-1}Xn,n$。实作应提升到 4×4 再求逆，不能把 3×4 矩阵的“逆”当作普通方阵求逆。

网络 F(I1,I2) 输出 X1,1 和 X2,1。后者的 z 是第二图像点在第一相机中的坐标，不是第二相机自身的深度。要获得 D2，需把点变到相机 2，或反向运行 F(I2,I1) 得到 X2,2。这个接口细节直接影响点云融合、遮挡检测和机器人抓取高度。

可复算例：焦距 500 px，主点 (320,240)，像素 (420,240)，轴向深度 2 m，得到点 (0.4,0,2) m，range≈2.0396 m。此差异虽只有约 4 cm，却已经远大于许多接触任务的容差。

$$\mathrm{Xcam}(u,v) = Z\,K^{-1}[u,v,1]^{T}$$
$$Xn,m = Tm,w\,(Tn,w)^{-1}\,h(Xn,n)$$
$$D_1 = z(X^{1,1});\quad D_2 = z(X^{2,2}),\ \text{not}\ z(X^{2,1})$$

量	含义	算例
Z	轴向深度	2 m
X	横向位置	(420−320)×2/500=0.4 m
range	到光心距离	√4.16≈2.0396 m

Depth is axial z, not Euclidean range. Back-project with inverse intrinsics. At a 100-pixel horizontal offset, a 500-pixel focal length and two-meter z depth give (0.4,0,2) meters and range about 2.0396 meters.

Both forward outputs share camera 1 coordinates. The second pointmap's z channel is therefore not camera 2 depth. Transform it or infer the reversed pair.

Use homogeneous four-by-four transforms. Compact pose notation in a paper should not become an ordinary inverse of a three-by-four implementation array.

03 / 双分支与交叉注意：共享信息，不共享坐标错误

A03 · 原文 PDF p.4,6,17 · 固定版本溯源；本页含原创分析

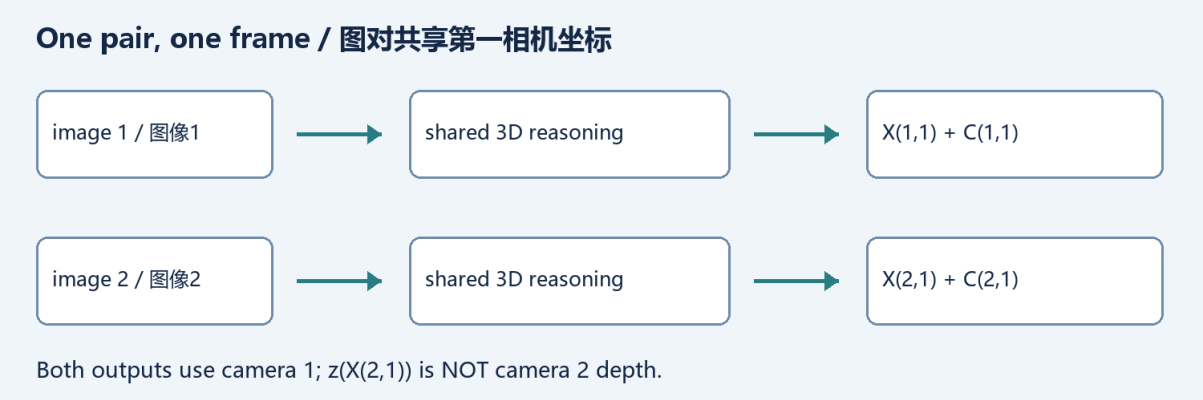
两张图先通过共享权重的 ViT 编码器形成 token，然后送入两个会持续交换信息的 Transformer 解码器，最后由回归头输出点图和置信图。共享编码器对两图采用同类视觉表征；交叉注意使每一分支能利用另一张图的纹理、遮挡和视差信息。

第一分支输出自身相机系的点，第二分支输出同一参考系的点。两个回归头的任务并不对称；简单把图像顺序交换而仍按旧参考系解释输出，会把相机、点和尺度全部混在一起。数据接口应把 reference_view_id 与 pointmap 一起保存。

原作默认采用 ViT-Large 编码器、ViT-Base 解码器，最终训练阶段使用 DPT 头，并利用 CroCo 预训练。附录给出先 224 低分辨率、再 512 高分辨率线性头、最后 DPT 的分阶段训练；这与每个场景从零优化 NeRF 权重不同。

注意力提供跨图信息通路，但不是显式多视几何定理。反光、透明、缺少重叠或训练域外纹理仍可让网络靠单目先验猜测。置信图和多图一致性应作为筛选线索，而不是直接当成经过标定的错误概率。

$$F_1=\mathrm{Encoder}(I_1);\quad F_2=\mathrm{Encoder}(I_2)$$
$$G_1 \leftarrow \mathrm{Decoder}_1(G_1,G_2);\quad G_2 \leftarrow \mathrm{Decoder}_2(G_2,G_1)$$
$$\mathrm{Head}_1 \rightarrow (X^{1,1},C^{1,1});\quad \mathrm{Head}_2 \rightarrow (X^{2,1},C^{2,1})$$



原创接口示意；两幅点图处于第一相机参考系。

A shared ViT embeds both images. Two decoders repeatedly exchange information through cross-attention; separate heads predict pointmaps and confidence. Outputs are asymmetric because both use the first view as reference.

Store reference-view identity with each pointmap. Reversing a pair changes coordinate interpretation. The default combines a ViT-Large encoder, ViT-Base decoder and a final DPT head after progressive training initialized from CroCo.

Cross-attention is an information pathway, not a projective-consistency guarantee. Low overlap, transparent surfaces and unfamiliar appearance can shift predictions toward monocular priors. Confidence requires independent calibration for safety-critical use.

04 / 尺度归一化回归：知道什么、不知道什么

A04 · 原文 PDF p.4,5 · 固定版本溯源；本页含原创分析

预测点图和真值分别除以各自有效点的平均到原点距离，再计算三维欧氏残差。这样训练目标不强迫模型从 RGB 恢复不可观测的绝对尺度，而是对形状及相对布局学习更直接的监督。有效 mask 决定哪些点参与尺度估计，必须同时用于归一化和损失，不能把天空或无效深度当成零点混入。

设真值点为 (1,0,0)、(0,2,0)，预测正好为其两倍。真值归一化尺度为 1.5，预测为 3，归一化后两者完全相同，损失为 0。这说明零训练残差并不意味着预测已经米制；正尺度倍数是这个目标有意忽略的自由度。

归一化也使不同点的梯度耦合：改变任意点会改变整组尺度 z，因此它不仅影响自己的残差。若实现者为了方便把 z 从计算图 detach，优化目标会与完整可微归一化不同；应把这一决定写入训练约定，而不是把不同版本默认为同一种算法。

点图的规范固定在第一相机，但尺度仍不确定。要让抓取器使用，需要至少一个米制锚与相机到机器人坐标变换。尺度因子及其不确定性应与点图一起传递，不能只在可视化时手动“调到看起来合适”。

$$z = [\sum v \sum i \in Dv \ \|Xv,i\|] \ / \ [|D_1| + |D_2|]$$
$$\ell v,i = \|Xv,i/z - Xgt,v,i/zgt\|_2$$
$$Xpred = sXgt,\ s>0 \Rightarrow \ell v,i = 0$$

归一化项	真值	两倍预测
有效点模长	1, 2	2, 4
平均尺度	1.5	3
归一化模长	2/3, 4/3	2/3, 4/3

Regression independently normalizes predicted and target pointmaps by mean valid-point norm. This removes positive global scale while retaining relative shape and layout.

For targets (1,0,0) and (0,2,0), doubled predictions change normalization from 1.5 to 3 and leave normalized residuals zero. This is a designed gauge, not recovered metric scale.

Normalization couples gradients through the shared scale; detaching it changes optimization. A robot needs a metric anchor, its uncertainty and a camera-to-robot transform as part of the geometry interface.

05 / 置信损失：权重为何不会全部归零

A05 · 原文 PDF p.5 · 固定版本溯源；本页含原创分析

置信加权损失由 Cℓ−αlogC 构成。第一项鼓励把高误差点权重降低，第二项惩罚置信度过小；原作把 C 参数化为 1+exp(g)，保证 C>1。因此这里的“置信度”不是 [0,1] 概率，也不能把数值 10 解释成九成正确。

固定几何残差 ℓ>0，若先忽略 C>1 约束，令对 C 的导数为零，可得 C*=α/ℓ；二阶导数 α/C²>0，故这是极小值。加上下界后，最优值趋近 max(1,α/ℓ)。如果几何残差趋于零，单点理想最优置信度会变得很大；实际网络容量、数据和参数化限制共同决定数值范围。

例如取教学参数 α=0.2，残差 0.05 的无约束最优 C=4，残差 0.5 的无约束最优 C=0.4，但参数化只能逼近下界 1。此时两点并非完全抛弃关系：难点仍以至少接近 1 的权重参与训练，促使网络在单视图区域外推。

工程上可按 C 过滤低质量点，但阈值应在目标域用独立误差标定，且检查被删掉的区域是否总是细杆、反光或透明边缘。平均误差下降可能只是覆盖率下降；应同时报告保留比例和错误分布。

$$\mathrm{Lconf} = \sum(C\ell - \alpha \log C);\quad C=1+\exp(g)>1$$
$$\partial L/\partial C = \ell-\alpha/C;\quad \partial^2 L/\partial C^2=\alpha/C^2>0$$
$$C^* \approx \max(1, \alpha/\ell) \text{ for fixed } \ell>0$$

教学例 α=0.2	ℓ=0.05	ℓ=0.5
无约束 C*	4	0.4
有下界 C>1	4 可达到	趋近 1
含义	较高权重	仍参与训练

Confidence balances residual weighting against a logarithmic penalty. Since C=1+exp(g)>1, these values are not probabilities.

For fixed positive residual, the unconstrained optimum is alpha divided by residual. A lower bound clips this toward one for difficult points instead of discarding them. As residual tends to zero, ideal confidence can grow without bound.

The example's alpha of 0.2 is illustrative, not a paper setting. Calibrate deployment thresholds using independently measured error and retained coverage. Filtering may delete exactly the fine reflective structures a robot needs to detect.

06 / 从点图找对应：互近邻不是身份真值

A06 · 原文 PDF p.5,7 · 固定版本溯源；本页含原创分析

因为两幅点图已在同一参考系，可直接在三维坐标中寻找最近邻。为了减少单向歧义，原作保留互为最近邻的像素对，即 i 的最近点是 j、j 的最近点也回到 i。这比单向近邻更保守，但仍依赖预测坐标的正确性、覆盖和分辨率。

最小反例：第一图的两个点在 0 和 0.1，第二图只有一个点在 0.04。第一图两点都会选这个点，但反向近邻只回到 0；互近邻只保留一个对应，从而减少多对一。然而若网络把错误物体预测到同一个位置，互近邻仍然可能稳定地保留错误关系。

对于遮挡边界，像素对之间可能本来就没有共同可见三维点，不应为了覆盖而强行匹配。对于周期纹理，坐标场可能整体偏移一个周期；增加局部置信阈值也无法证明匹配身份。需要结合循环一致性、图像可见性和已知相机模型作核验。

流体粒子追踪还多了时间身份要求：共享三维位置的近邻只回答“哪个点最接近”，不回答“是否同一粒子”。粒子密度高、进出光片或交叉运动时，应使用运动门控、质量守恒或多帧轨迹证据，且把未确定关联保持为不确定状态。

$$\mathrm{NN}_2(i) = \operatorname{argmin}_j \|X^{1,1}i - X^{2,1}j\|_2$$
$$M = \{(i,j): j=\mathrm{NN}_2(i) \text{ and } i=\mathrm{NN}_1(j)\}$$
$$\text{reciprocal nearest neighbor} \neq \text{verified temporal identity}$$

风险	近邻会发生什么	可增加的证据
遮挡	强配到邻近背景	可见性/深度次序
周期纹理	稳定但错误偏移	循环及重投影
粒子交叉	身份互换	多帧动力学/门控

Reciprocal nearest neighbors reject a candidate unless each point chooses the other. This reduces many-to-one matches without proving identity.

If one map has positions 0 and 0.1 and the other has 0.04, both forward searches choose the same point; reverse matching retains only the first. A consistently wrong location can still pass.

Occlusions may have no valid counterpart; repeated textures can produce coherent shifts. Particle tracking also requires temporal identity. Combine visibility, cycles and motion evidence rather than forcing every pixel into a track.

07 / 恢复焦距：可解析的教学近似

A07 · 原文 PDF p.5,16 · 固定版本溯源；本页含原创分析

原作在主点近似位于图像中心、方形像素的假设下，从第一相机系点图估计焦距。令中心化像素位置为 a=(u−cx,v−cy)，三维点的归一化平面坐标为 b=(X/Z,Y/Z)，则理想模型给 a=fb。原文采用置信加权向量残差的范数，可通过 Weiszfeld 类迭代求解。

为了看清可观测性，本册另推导平方残差的教学版本：最小化 Σwi||ai−fbi||²，对 f 求导，得到标量加权最小二乘解。它不是把原论文的非平方目标偷偷替换掉；两者的离群点响应不同，正式实现应选择并记录目标。

取两点 b=(0.2,0)、(0,0.1)，像素偏移分别为 (100,0)、(0,50)，等权平方解恰好得到 f=500 px。如果所有点都非常接近主轴，分母 Σwi||bi||² 接近零，焦距会高度不稳定；若主点偏移或径向畸变未建模，残差会被焦距或点图形状吸收。

应优先使用可信标定内参来约束最终米制流程，再把预测焦距视作无标定入口或失配诊断。对第二相机必须使用自身坐标点图，而不能把 X2,1 当成 X2,2 代入。

$$\text{paper: } \min f \sum_i C_i \, \|a_i - f b_i\|_2$$

$$\text{teaching LS: } f_{LS} = \left[\sum_i w_i \, (a_i \cdot b_i)\right] / \left[\sum_i w_i \, \|b_i\|^2\right]$$

$$(100\times0.2+50\times0.1)/(0.2^2+0.1^2) = 500 \text{ px}$$

假设	若不成立	后续处理
中心主点	系统像素偏移	联合估计/使用标定
方形像素	横纵焦距不同	引入 fx、fy
针孔无畸变	边缘点曲线误差	先去畸变/扩大模型

With a centered principal point and square pixels, normalized point coordinates b map to centered pixel coordinates a through focal f. The paper minimizes confidence-weighted vector norms using an iterative solver.

A squared-residual teaching variant has a closed-form weighted LS solution, with different outlier sensitivity. Normalized offsets 0.2 and 0.1 paired with pixel offsets 100 and 50 recover 500 pixels.

Near-axis points offer little focal information. Principal-point errors and distortion can be absorbed into focal or shape. Use calibration when available and ensure the pointmap is in the fitted camera's own frame.

08 / 从相似变换到机器人坐标

A08 · 原文 PDF p.5,7,15,16 · 固定版本溯源；本页含原创分析

对两个包含共同三维点、但处于不同坐标系的点集，可以估计旋转、平移及相对尺度。带权 Procrustes 先求加权中心，中心化后组成互协方差矩阵，再用 SVD 求正交旋转；通过行列式修正避免把反射误当旋转，最后求尺度和平移。

这个闭式对齐要求对应关系正确并且几何不退化。重合点或近共线点会使部分旋转不稳定，离群点还会显著拉偏结果。原文因此也提出更稳健的 PnP-RANSAC 路线：把点图产生的 2D—3D 对应与相机模型结合，利用采样共识拒绝离群点。

尺度只能从一个点集转移到另一个点集，不能从相似变换本身凭空产生米制。若参考地图本身无尺度，则相对位姿仍然无米制平移长度。若参考点稀疏且分布不良，尺度估计会不稳定；附录无已知焦距定位在 Cambridge 上误差变大，作者把它关联到数据库点图太稀疏而难以可靠缩放。

具身接口应继续乘上相机到基座的手眼变换，并传播尺度、姿态和点误差。只有地图点坐标、相机变换与机器人碰撞模型都处于同一单位和时间基准，才可比较末端目标和障碍净空。

$$\min_{s,R,t} \sum_i w_i \|y_i-(sRx_i+t)\|^2,\quad R^TR=I,\ \det R=1$$
$$H=\sum_i w_i (x_i-\bar{x})(y_i-\bar{y})^T;\quad H=U\Sigma V^T$$
$$R=V\,\mathrm{diag}(1,1,\det(VU^T))\,U^T;\quad t=\bar{y}-sR\bar{x}$$

坐标链	变换	保持项
点图→相机	相似/刚体配准	尺度来源
相机→基座	手眼外参	时间与单位
基座→工具	正运动学	工具实际几何

Weighted Procrustes estimates relative rotation, translation and scale. Center correspondences, compute cross-covariance and apply SVD with a determinant correction to exclude reflection. Then recover scale and translation.

Correct matches and nondegenerate geometry are essential. Outliers and collinearity destabilize alignment; PnP-RANSAC offers a robust camera-estimation route.

Similarity alignment transfers reference scale, not creates metric information. Sparse reference pointmaps hinder the appendix's unknown-focal localization. A robot additionally needs hand-eye calibration and uncertainty propagation before comparing reconstructed points with collision geometry.

09 / 多视图全局对齐：图连通与规范约束

A09 · 原文 PDF p.5,6,14 · 固定版本溯源；本页含原创分析

每对图像的局部点图都使用不同参考系与未知尺度。DUSt3R 先建立图：图像为节点，有共同视觉内容的图对为边。每条边引入相似变换，把两幅局部点图一起变到公共世界；同一图像在多条边产生的预测通过全局点图 χ 相互约束。

原作目标是置信加权的三维欧氏距离，并约束所有边尺度的乘积为 1，防止把尺度同时缩到零。数值实现可令边尺度 σe=exp(ae)，再对 ae 去均值，从而精确实现乘积为 1。仍需固定整体世界参考系；尺度约束不会自动消除全局旋转和平移自由度。

如果图不连通，每个连通分量都能独立移动，不能宣称它们已在一个可比坐标系。即便连通，桥接边若是错误匹配，也可能把两个局部正确的子图拼成错误场景。应报告连通分量、边残差与闭环一致性，必要时拒绝不可靠桥边。

24 张图的完全图有 276 对，4 张图只有 6 对。稀疏图策略能减少二次增长，但必须保证有效覆盖而不仅是“边数够”。这些组合数是解析计算，不是本机 DUSt3R 推理耗时。原作报告的 GPU 速度不能被移作 CPU 延迟承诺。

$$\mathrm{LGA} = \sum_{e}\sum_{v\in e}\sum_{i} C_{v,e,i}\,\left\|\chi_{v,i} - \sigma_e\, T_e\, h(X_{v,e,i})\right\|_2$$

$$\prod_e \sigma_e = 1;\quad \sigma_e = \exp(a_e - \mathrm{mean}(a))$$

$$\text{complete graph edges} = N(N-1)/2$$

图性质	工程意义	检查
不连通	分量间位置无约束	报告分量/拒绝强拼
弱桥边	可把子图接错	残差与循环一致性
全连接	成本二次增长	检索+覆盖选边

Global alignment connects pairwise predictions through an image graph. Each edge shares a transform and scale between its two pointmaps; global maps reconcile repeated observations.

A product-of-scales constraint prevents collapse. Zero-mean log scales enforce it, but the overall rigid gauge still needs an anchor. Disconnected components have no established relative placement.

Inspect bridges, residuals and cycles. Complete pairing grows from six edges at four views to 276 at 24 views; these are combinatorial counts, not measured inference times. Sparse pairing should preserve genuine overlap and connectivity.

10 / 读懂原作数字：深度对齐与重建精度

A10 · 原文 PDF p.7,8,9,15,16 · 固定版本溯源；本页含原创分析

原作多视图深度表中，DUSt3R 不使用已知位姿、深度范围和内参，但评价前对预测与真值作中位数尺度对齐。512 模型在 ETH3D 的相对误差为 2.91%，五域平均为 4.73%。这些数字说明形状和相对深度能力，不能直接写作“无需尺度锚即可获得 4.73% 米制深度”。

DTU 的点云重建使用真值坐标对齐，报告 accuracy 2.677 mm、completeness 0.805 mm、overall 1.741 mm。accuracy 是预测点到真值的最近距离，completeness 是真值点到预测的最近距离；它们都是定向距离的均值，不是每个点最大误差，也不是全场景安全误差上界。

同表 GeoMVSNet overall 为 0.295 mm，并使用已知相机且有域内训练。DUSt3R 的便利性与泛化接口很强，却不是在该精细表面指标上全面超越标定 MVS。选择模型应依据输入条件和任务容差，不要把不同先验条件的数值排成同一公平速度榜。

姿态评价也分旋转、平移方向和组合角度准确率。10 帧 CO3Dv2 中，PnP 路线 mAA(30)=77.2，全局对齐为 76.7；RealEstate10K 则分别为 61.2 与 67.7。下游路线存在反转，证明“统一基座”不意味着每个读出方式都同样最优。

$$\text{accuracy} = \text{mean}_{p\in P}\ \min_{q\in Q} \|p-q\|$$
$$\text{completeness} = \text{mean}_{q\in Q}\ \min_{p\in P} \|q-p\|$$
$$\text{overall} = (\text{accuracy}+\text{completeness})/2 = 1.741\ \text{mm}$$

原文表	条件	数值
Table 3 ETH3D rel	中位数对齐	2.91%
Table 3 五域平均 rel	中位数对齐	4.73%
Table 4 DTU overall	对齐真值坐标	1.741 mm
Table 4 GeoMVSNet	已知相机/域内训练	0.295 mm

Multiview-depth evaluation aligns median predicted and reference depths. Its errors describe aligned shape, not unanchored metric sensing: the 512 model reports 2.91 percent on ETH3D and a 4.73 percent five-domain mean.

Aligned DTU reconstruction gives directed mean distances of 2.677 millimeters for accuracy and 0.805 for completeness, averaging 1.741. These are not maximum-error guarantees. Calibrated, domain-trained MVS has lower errors under different conditions.

Readout matters: PnP slightly leads global alignment on CO3Dv2 mAA, while the order reverses on RealEstate10K. Choose a route using target constraints rather than assuming all outputs are equally strong.

11 / 训练与失败模式：先验的来源必须可见

A11 · 原文 PDF p.3,4,6,9,14,16,17 · 固定版本溯源；本页含原创分析

DUSt3R 使用约 8.5M 图像对的混合训练，并在每个 epoch 中平衡不同数据集采样。CroCo 预训练与较高分辨率在多数下游表中带来持续收益，但模型仍是在已见分布上学会推断形状和相机规律，不能从“支持任意图集”推导出“对任意材质均正确”。

透明面违反每像素单个表面点的基本近似，天空没有有限可靠深度，强反射让外观随视角剧变，动态物体则破坏静态多视图一致性。置信图可提示困难，但高置信系统性偏差仍可能发生。要用独立几何与目标域样本检验置信—误差关系。

附录中心裁剪相当于焦距增强，并维持训练中的主点居中；这解释了恢复焦距的简化假设有数据支持，也说明大主点偏移、鱼眼或未经去畸变相机需要专门评估。原文同一相机不同帧的内参也可以独立优化，不要把它说成严格共享内参的标定结果。

推荐保存 raw pointmap、confidence、有效 mask、原始/缩放图像尺寸、参考视图、尺度锚与对齐残差。只导出一份“漂亮点云”会丢失检查错误对应与坐标转换所需的索引，后续修复也难以回到原始观测。

$$\text{pair metadata} = \text{image IDs} + \text{resize} + \text{reference frame} + \text{mask}$$
$$\text{geometry metadata} = \text{scale anchor} + \text{transform} + \text{residual} + \text{uncertainty}$$

失败类	表现	对应验证
透明/天空	不存在表面的预测点	材质掩码/独立深度
动态	重复或拉伸	同步拍摄/动态分割
域外相机	边缘几何与焦距偏移	标定与去畸变
低重叠	单目猜测主导	覆盖与额外视角

Mixed image-pair training, dataset balancing and CroCo initialization supply learned geometric priors. Unconstrained input support does not imply universal material accuracy.

Transparency, sky, reflections and motion challenge single-surface static assumptions. Centered training crops also identify off-center or distorted cameras as important test cases. Confidence can be systematically wrong and requires target-domain checks.

Retain pixel-indexed geometry, confidence, masks, resize metadata, reference identity, scale anchors and residuals. A visually attractive exported cloud discards information needed to diagnose transformations and recover the connection to measurements.

12 / 流体与具身的优先迁移：几何骨架加物理观测

A12 · 原文 PDF p.3,4,5,6,9,16 · 固定版本溯源；本页含原创分析

推荐把 DUSt3R 用作场景几何骨架：先恢复实验容器、固定壁面和机器人工作区，再用米制标靶固定尺度，最后把稀疏粒子或事件观测放到该几何里。这样把容易静态重建的物体与高速流场分开，而不是要求一个 RGB 几何网络同时解决透明介质、粒子身份和不可压缩速度。

以下为待验证设计：DUSt3R 提供候选相机与壁面，事件视觉提供高时间分辨率亮度约束，流匹配模型在满足边界的速度/轨迹空间生成候选，散度投影与最优传输分别约束质量连续性和分布关联。各模块必须保留误差与支持区，不能把先验补全写成直接测量。

最小判别实验可固定容器与标定，以独立测得的壁面和少量米制点作参照，比较纯点图、加标定重投影、加物理约束三组。在同一相机覆盖下分别报告壁面位置误差、粒子重投影、速度误差、边界通量与未知区域；再用增加视角测试哪些误差真因可观测性改善而降低。

本册附 CPU 解析测试：反投影、尺度不变性、置信最优值、焦距平方近似和图规模。原文全部 23 页保留逐页提取与版面块坐标，相应页号映射在 source_map。所有新数值例标为教学算例，未运行网络、不下载模型权重、不宣称本地推理性能。

> world metric geometry → visibility / boundaries / camera rays
> observations + flow prior → candidates + uncertainty
> independent measurements → discriminate, calibrate, accept

术语	固定中文	含义
pointmap	点图	逐像素三维坐标场
global alignment	全局对齐	融合图对参考系
confidence	置信分数	正权重，非概率
scale gauge	尺度规范自由度	RGB 相同的缩放自由度
reciprocal matching	互近邻匹配	双向最近邻筛选

Use DUSt3R as a static workspace skeleton, fix metric scale with targets and treat fast fluid measurements separately. Container walls and cameras are different inference objects from transparent flow and particle identity.

The proposed hybrid combines geometry, event-based timing, flow candidates and explicit physical projection or transport. Preserve uncertainty and observational support across interfaces.

Compare raw pointmaps, calibrated reprojection and physical constraints against independent metric references under identical coverage. Measure geometry and flow, not just appearance. Included CPU tests establish analytical premises; no network inference or timing benchmark is claimed.