单目深度基础模型：数据、几何与具身接口

01 / 深度先验很强，几何承诺要分层

arXiv:2406.09414v2，pp.1–3；本册独立问题定义

Depth Anything V2 是判别式单目深度基础模型：给一张 RGB，预测每像素的深浅结构。它的关键路线是用精确合成深度训练大教师，再用教师给真实图像打伪标签，最后让学生在真实图像分布上学习。这样把“监督是否精确”和“图像是否多样”拆开处理，而不是把所有带标签的数据一股脑混合。

本册首先固定接口含义。普通相对深度模型输出的是允许尺度和平移变化的逆深度分数，不是默认以米为单位的 Z；模型专门微调后的 metric 分支则是另一项产品与训练契约。用于机器人时，深度边界清晰、点对排序正确、世界系定位准确和安全净空可靠是四个不同等级的承诺，不能从前两项自动跳到后两项。

底本为 2024-10-20 的 arXiv:2406.09414v2，共 30 页，NeurIPS 2024。作者为 Lihe Yang、Bingyi Kang、Zilong Huang、Zhen Zhao、Xiaogang Xu、Jiashi Feng、Hengshuang Zhao。论文 arXiv 默认 non-exclusive 分发许可，与 GitHub 代码及各尺寸权重许可相互独立；本册因此采用原创推导和工程分析，不制作逐段替代译本。

$$R = f_\theta(I), \quad I \in \mathbb{R}^{B\times 3\times H\times W}, \quad R \in \mathbb{R}^{B\times H\times W}$$

$$\rho = aR + b, \qquad Z = 1/\rho$$

输出 / Output	能直接回答	仍需条件
相对逆深度	哪个点更近	排序方向/有效掩码
metric 分支	特定训练域的米制深度	域外误差/尺度核验
世界系点云	三维空间位置	内外参/时序/有效深度
安全净空	轨迹是否避障	误差界/工具几何/完整覆盖

> A strong monocular depth prior does not directly provide metric world geometry. The standard model predicts affine-invariant inverse depth; metric checkpoints have a separate training and usage contract. Fine boundaries, correct pairwise ordering, accurate coordinates and certified clearance are distinct claims. The useful engineering question is which of these the downstream interface actually requires, and which measurements can establish it.

02 / 数据设计：准确标签与真实分布不是同一轴

原文 pp.3–6 §2–5；本册数据偏差分解

真实深度标签可来自传感器、双目匹配或 SfM，错误机制各异：透明物体可能没有有效读数，重复纹理破坏匹配，动态区域不服从静态多视图假设。监督中缺失的细孔、椅腿或透明边界不是靠把损失权重调大就能凭空补齐。合成图像可以提供这些几何细节，但其外观和场景覆盖又不等于真实世界。

将标签误差写成独立零均值噪声通常过于乐观。若错深度总出现在玻璃、细结构和反光上，它与输入语义相关，最小风险训练会把这种偏差学成稳定规则。增加同类样本只会让错误学得更确定。相反，合成域到真实域的差异出现在输入分布上；即便监督完全精确，也不能保证模型在没见过的真实布局上泛化。

原文采用约 595K 合成图像训练教师，再借助约 62M 无标签真实图像进行标签级蒸馏。重要的不是“伪标签必然优于人工标签”，而是大教师的有效先验、精细几何监督和更合适的输入分布在这个配置中形成互补。工程上应分别审计标签偏差、场景覆盖和模型容量，不把三项揉成一个“数据规模”数字。

$$y_{\text{obs}} = y^{*} + b(I) + \epsilon, \quad \mathbb{E}[\epsilon \mid I] = 0$$

$$\mathbb{E}[y_{\text{obs}} \mid I] = y^{*} + b(I)$$

> Label quality and input coverage occupy different axes. Structured supervision errors are not averaged away by simply adding more similarly biased labels. Synthetic data improve geometric completeness but introduce a domain gap. Label-level distillation exposes students to real inputs while retaining the teacher's detailed predictions. Its success depends on teacher quality and coverage; “pseudo” is not itself a quality guarantee.

独立分析 / Independent analysis：b(I) 用于解释结构性标签偏差，不是论文额外训练的网络。

03 / 三阶段教师—学生机制

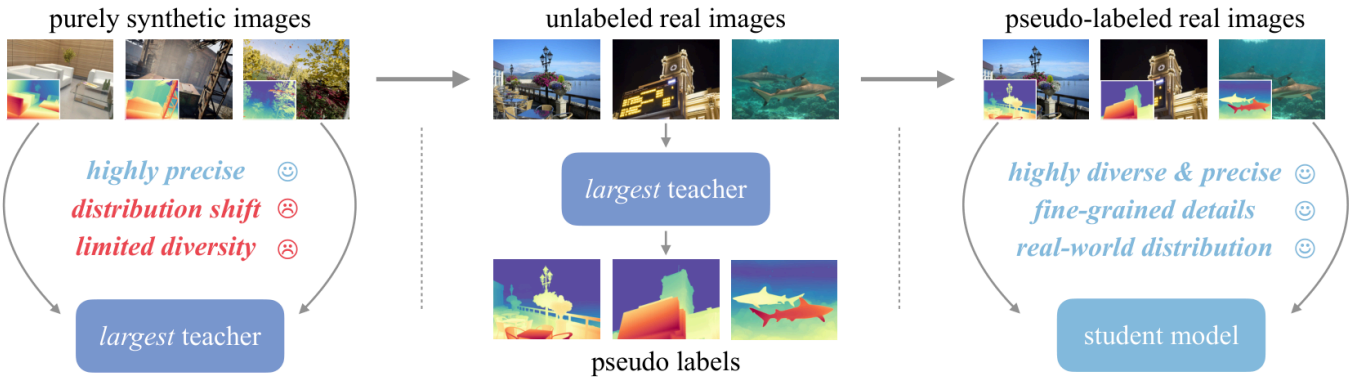
原文 p.6 Figure 7、§5；本册训练契约分解

第一阶段的教师以 DINOv2-G 为编码器，在精确合成监督下学习深度。第二阶段冻结教师，给真实图像产生伪深度；这一步把教师的场景先验变成可被不同尺寸学生学习的预测目标。第三阶段训练学生，最终方案主要使用伪标注真实图像，而不是强行把合成与真实标签混在同一学生批次。

这与让小模型直接在合成域学习不同：学生不再独自承担合成到真实的跨域外推。它看到的是最终部署更接近的真实外观。与此同时，教师并不能在完全未知区域凭空保证正确；学生可能复制教师偏差，也可能借助真实数据的规律改善泛化。要区分“学生拟合教师”与“学生接近真值”，必须保留独立标注测试。

原文报告 518×518 训练裁剪；教师 batch 64、160K 次，学生 batch 192、480K 次，编码器和解码器学习率分别为 5×10^{-6} 和 5×10^{-5}。它们是论文训练设置，不是本机已执行的开销。对于本轮 CPU 文献工作，我们只审计模型与证据，不下载权重、不重训，也不把原作训练预算写成本地性能。

```
1. train teacher on synthetic RGB/depth  # 精确监督学几何
2. freeze teacher; infer on real RGB      # 固定教师生成标签
3. trim top 10% pseudo-label loss regions  # 降低潜在噪声影响
4. train student + feature alignment      # 真实域压缩知识
5. test on independent labeled images     # 教师一致≠真值正确
```



Yang et al., Figure 7，p.6。原作训练桥接机制，300 DPI 原图裁图。

> Stage one learns from precise synthetic depth. Stage two creates a frozen teacher's predictions on real images. Stage three trains deployable students on those real inputs. This removes the need for each small student to solve synthetic-to-real transfer alone. Teacher agreement is still different from correctness, so an independent evaluation remains essential. Published training settings describe the original study, not a local rerun.

04 / 尺度与平移不变：先说清丢掉了什么

原文 p.6 §5.2；本册独立不变性推导

相对逆深度容许正比例缩放和平移：两个预测只要相差 aR+b，仍可表示相同的前后结构。这样的自由度适合汇集不同来源的数据，却使输出不能直接接入要求米制距离的碰撞检测。单幅图中的尺度歧义并不会因为模型参数增大而在数学上消失。场景先验能给出合理猜测，但不等于标定证据。

一种常见的 SSI 构造先减去中位数，再除以平均绝对偏差，然后在有效像素上比较归一化后的预测与标签。对正比例仿射变换，归一化形式保持不变；若尺度趋近零，分母需要明确的数值处理。本节展示 MiDaS 风格的概念构造以解释论文所用 SSI 类目标，不声称它是未公开训练实现逐行相同的唯一公式。

另一种用于工程校准的办法是在少量稀疏 metric 逆深度锚点上求最小二乘 a,b。设计矩阵的两个列分别为预测值和常数；只有锚点有足够深度变化，矩阵才满秩。所有锚点近似同深度时，虽然像素很多，仍不能稳定辨识尺度与偏移。要报告拟合残差与条件性，而不是只报告锚点数量。

$$\widetilde{R} = \frac{R - \operatorname{median}(R)}{\frac{1}{M}\sum_i |R_i - \operatorname{median}(R)|}$$

$$\min_{a,b} \sum_i w_i (a R_i + b - \rho_i)^2, \quad \rho_i = 1/Z_i$$

$$\begin{bmatrix} \sum w_i R_i^2 & \sum w_i R_i \\ \sum w_i R_i & \sum w_i \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum w_i R_i \rho_i \\ \sum w_i \rho_i \end{bmatrix}$$

独立校准例	锚点 1	锚点 2	锚点 3
Raw inverse score R	0.2	0.7	1.9
Measured inverse depth ρ (m^{-1})	0.18	0.38	0.86
Recovered affine map	a=0.4	b=0.1	ρ=0.4R+0.1

> Affine-invariant training deliberately discards two degrees of freedom. A median/MAD-style normalization illustrates the invariance; sparse metric calibration instead solves for the missing affine parameters. These are different operations. Many anchors on one plane at nearly the same depth may leave the fit ill-conditioned. Increasing model capacity does not remove the geometric scale ambiguity of monocular input.

05 / 梯度匹配为何需要真实的细节监督

原文 p.6、p.14 §B.7/Figure 10；本册损失机制解读

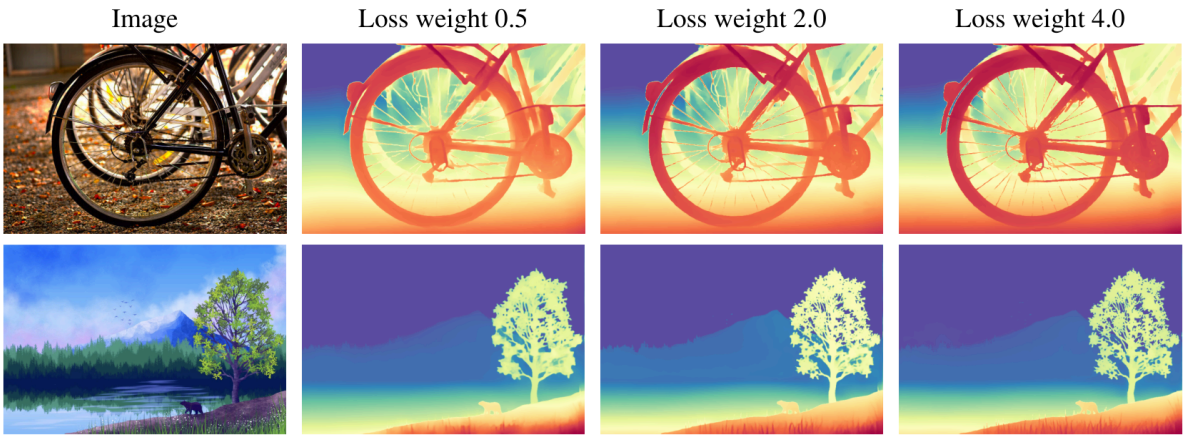
深度平滑并不总是表示模型欠拟合：如果标签在细结构处被抹平，梯度匹配会明确奖励抹平。对残差 R−Y 求空间差分，再最小化其绝对值，可以逼近边缘，但不能告诉模型被标签漏掉的椅腿应该在哪里。梯度项的收益因此取决于监督的空间频率是否可信。

原文在精确合成标签上观察到增大梯度匹配权重有助于细节；默认最终 SSI:GM 权重为 1:2，附录展示 GM 从 0.5 到 4.0 的视觉变化，并选择 2.0 平衡细节和指标。图像更锐不意味着所有传统误差必然单调改善；薄结构只占少量像素，平均指标可能对它不敏感。

伪标签训练时，原作每个样本忽略损失最高的 10% 区域，并附加特征对齐以保留 DINOv2 的语义信息。这个鲁棒策略减少可能的错伪标签影响，也会降低某些真正困难像素的训练权重。工程上应检查被丢弃像素是否集中在接触边界、透明物体和工具尖端，避免以全图损失下降换来控制最关键区域的缺失。

$$\mathcal{L}_{\text{gm}} \propto \sum_{s,i} \left(|\partial_x(\widetilde{R} - \widetilde{Y})| + |\partial_y(\widetilde{R} - \widetilde{Y})| \right)$$

$$\mathcal{L} = \mathcal{L}_{\text{ssi}} + 2\mathcal{L}_{\text{gm}} + \lambda_f \mathcal{L}_{\text{feature}}$$



原图 10，p.14：0.5 / 2.0 / 4.0 梯度权重的定性对照。公式为本册机制表示，不补造未公开的特征损失权重。

> A gradient penalty can only preserve spatial detail that the target actually contains. With coarse labels it may reinforce coarse predictions. The paper selects a gradient weight of 2.0 after studying 0.5 through 4.0. Dropping the highest-loss 10% of pseudo-label regions improves robustness but may also exclude important difficult regions. Inspect the rejection mask rather than treating the trimmed objective as neutral.

06 / 网络张量维度：ViT 不直接输出点云

原文 p.8 §7.1；官方 depth_anything_v2/dpt.py；本册形状推演

以方形 518 输入为例，DINOv2 的 patch 边长为 14，得到 37×37=1369 个空间 token。编码器提取四个中间层，而不是把最后四层误当多分辨率输入。以 Small 为例，每层空间特征约为 [B,1369,384]；还可带独立 class token，但默认解码并不依赖 class-token 读出。

DPT 先把各层 token 重排回 [B,C,37,37]，投影到不同通道并做多尺度重组，再以特征融合块逐级合并。这里的“多尺度”主要由重组和上采样产生；四层 ViT 空间 token 本身在同一网格上。Small 的投影通道配置为 48、96、192、384；融合宽度 64。最终输出经非负激活并调整到输入/原始图像尺寸，得到一张标量深度分数图。

一般图片保持长宽比并使尺寸满足 patch 整除；不能强行把所有宽高都写成 518，否则几何投影时会使用错误像素坐标。若后续生成点云，应保存 resize、crop 与 padding 的坐标变换，恢复到与内参对应的网格。把可视化色图作为深度输入会丢失数值关系，必须使用浮点原始预测。

Encoder	Token channels C	DPT fusion width	Small/large distinction
ViT-S	384	64	24.8M params (official README)
ViT-B	768	128	97.5M params
ViT-L	1024	256	335.3M params
ViT-G	1536	384	~1.3B params

```
RGB [B,3,518,518]
  -> patch14 -> spatial tokens [B,1369,C]
  -> 4 intermediate features (Small C=384)
  -> reshape 4 x [B,C,37,37]
  -> channel projection / scale reassembly
  -> DPT fusion (Small width=64)
  -> nonnegative raw score [B,518,518]
  -> resize to original grid [B,H,W]
  -> optional metric calibration -> back-project [B,N,3]
```



Preserve aspect ratio and original camera-grid transforms; a raw score map is not metric XYZ.

本册原创 Small 分支维度流图。四个 ViT 层共享 37×37 网格；DPT 负责多尺度重组与融合。

> Four intermediate ViT layers share a patch grid; DPT reassembles them into multi-scale feature maps. At 518 square input, the grid contains 1369 spatial tokens. The final output is a scalar map, not a point cloud. Preserve resize and padding transforms before using camera intrinsics, and consume raw floating-point predictions rather than a color visualization.

07 / 消融表复盘：伪标注真实域的增量

原文 p.9 Table 5、p.10 Table 6；本册计算与批判

表 5 直接比较仅合成监督、合成加伪标注真实、仅伪标注真实三条学生路线。Small 的 DA-2K 点对准确率依次为 89.8%、94.1%、95.3%；从仅合成到仅伪标注提高 5.5 个百分点。对应错误率从 10.2% 降至 4.7%，相对下降约 53.9%。这两种差值回答不同问题，不能把 5.5 个百分点写成 53.9 个百分点。

同一表中 Large 的混合方案 97.3%，略高于仅伪标签的 97.1%；因此“去掉合成对所有尺寸都最好”不成立。原作最终选择统一路线，较小尺寸收益更明确。传统基准也不是全面单调提升，例如 Table 2 的 Large 在 Sintel 上 V2 AbsRel 0.487，而 V1 为 0.458，数值更差。不能用新基准优势隐藏这个反例。

表 6 中，DIML 原人工标签训练得到 DA-2K 80.2%，教师伪标签训练为 89.7%，显示该配置下伪标签更有效；它并不证明任何模型生成标签都比人工标注好。相关因果解释应同时保留标注机制、教师来源与测试集构成，尤其新基准本身由包括当前模型在内的模型分歧辅助挑选难例。

DA-2K accuracy	Synthetic only	Mixed	Pseudo-real only
Small	89.8%	94.1%	95.3%
Base	92.9%	96.7%	97.0%
Large	96.0%	97.3%	97.1%

Encoder	$\mathcal{D}^l$	$\mathcal{D}^u$	KITTI [24]		NYU-D [70]		Sintel [8]		ETH3D [62]		DIODE [76]		DA-2K
			AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	Acc (%)
ViT-S	✓		0.104	0.889	0.084	0.928	0.518	0.702	0.155	0.827	0.087	0.926	89.8
	✓	✓	0.085	0.928	0.054	0.971	**0.491**	**0.723**	0.143	0.849	0.074	0.941	94.1
		✓	**0.078**	**0.936**	**0.053**	**0.973**	0.500	0.718	**0.142**	**0.851**	**0.073**	**0.942**	**95.3**
ViT-B	✓		0.094	0.912	0.062	0.963	0.618	0.715	0.148	0.842	0.076	0.940	92.9
	✓	✓	0.080	0.938	**0.049**	**0.976**	0.515	0.732	**0.137**	**0.859**	**0.068**	**0.950**	96.7
		✓	**0.078**	**0.939**	**0.049**	**0.976**	**0.495**	**0.734**	**0.137**	0.858	**0.068**	**0.950**	**97.0**
ViT-L	✓		0.081	0.937	0.048	0.976	0.516	0.731	0.133	0.864	0.071	0.949	96.0
	✓	✓	0.075	**0.947**	**0.045**	**0.979**	0.542	0.741	**0.130**	**0.866**	**0.066**	**0.953**	**97.3**
		✓	**0.074**	0.946	**0.045**	**0.979**	**0.487**	**0.752**	0.131	0.865	**0.066**	0.952	97.1
Teacher model (ViT-G)			0.075	0.947	0.044	0.979	0.530	0.767	0.131	0.865	0.066	0.954	97.4

原作 Table 5，p.9。Dl 指精确合成标签，Du 指伪标注真实图像；本册保留原表以核对多个测试域。

> Small gains 5.5 percentage points on DA-2K from synthetic-only to pseudo-real-only training, equivalent to a 53.9% relative error-rate reduction. Large slightly favors the mixed route in the same table. Conventional benchmarks also contain counterexamples. Keep these qualifications alongside the gain: the result supports this data pipeline, not a universal rule that pseudo labels or removing synthetic samples always wins.

08 / DA-2K：稀疏排序不等于稠密米制误差

原文 pp.7–8 §6、pp.16–17 §C；本册指标分析

DA-2K 的 1K 图像、2K 像素对覆盖八类场景，判断哪个像素更近。原作让多个模型对候选点对投票，分歧样本交给人标，并补充人工寻找的困难点对；透明和物体场景不完全依赖自动分歧。原作还排除难判定或争议对，并对标注做复核。它扩充了传统室内和道路基准的视野，但稀疏点对不能衡量每像素误差，更不能建立毫米级安全界限。

给所有深度加一个错误的全局尺度，像素排序完全不变。因此排序 97% 和 metric 定位误差 3% 没有这样的直接对应。对平坦物体表面，局部斜率或薄边缘错得很明显，也可能不影响抽中的两个点。模型选择时应把排序测试看作快速分场景筛查，再用独立密集或稀疏 metric 锚点验证具体用途。

DA-2K 的困难样本由部分模型分歧产生，会提高对这组模型既有薄弱点的灵敏度；所有模型共同错且未被人工挑中的情况则可能漏检。三人复核能降低标注错误，不能消除选择分布的影响。用它评估本机相机时，应另建不依赖候选模型选样的设备/任务测试集，尤其保留接触表面和碰撞走廊。

$$\mathrm{Acc} = \frac{1}{K}\sum_{k} \mathbf{1}[\mathrm{sign}(R_{i_k} - R_{j_k}) = s_k]$$

$$Z_i \mapsto \gamma Z_i, \quad \gamma > 0 \quad \Rightarrow \quad \text{ordering unchanged}$$

评价	可以发现	不能单独证明
DA-2K 点对	跨场景前后关系错	米制尺度/稠密边界
AbsRel / RMSE	有效真值上的深度偏差	未标注区安全
边界/法向误差	薄结构与局部几何	完整闭环稳定
机器人闭环	任务域实际结果	其他相机/其他物体泛化

> Pairwise ranking is invariant to many metric distortions. A 97% ordering accuracy does not imply a 3% depth error, nor a safe collision margin. Disagreement-driven selection is useful for finding known failure modes but can miss shared mistakes. Treat DA-2K as a diverse screening tool, then evaluate task-specific metric geometry on independently selected observations.

09 / 从逆深度到世界系：误差会在远处放大

本册独立几何推导；原作区分 relative/metric 分支

相对逆深度经过校准得到 ρ 后，Z=1/ρ。其导数为 −1/ρ²=−Z²，因此固定逆深度误差在远处会变成平方级增长的深度误差。举例说，δρ=0.01 m⁻¹ 时，在 1 m 附近约为 10 mm，在 3 m 附近约为 90 mm。这是一阶近似；若误差令分母接近零，必须直接判为无效，而不是裁成一个“很远”的安全点。

针孔反投影为 p_c=ZK⁻¹[u,v,1]，世界系再通过 R_wc p_c+t_wc 转换。深度误差沿视线传播，像素误差随 Z/f 放大，外参角度误差又会随距离产生横向位移。若距离 1 m、外参旋转误差 1°，横向误差量级已经约 17.5 mm，足以超过不少接触任务容差。

因此，“单目深度细节好”不能替代外参校准。构建 SDF 时应把深度有效掩码、未知空间与已观测自由空间分开；遮挡后的区域不能因为没有点就被默认清空。最低限度做法是给近障表面保守膨胀，并在实际闭环中单独验证接触和净空，但膨胀半径必须与测得误差相匹配，不能凭一个统一常数保证安全。

$$\delta Z \approx -Z^2 \delta\rho, \qquad \mathbf{p}_c = Z K^{-1} [u, v, 1]^\top$$

$$\mathbf{p}_w = R_{wc}\mathbf{p}_c + \mathbf{t}_{wc}, \quad \delta\mathbf{p}_w \approx R_{wc}\delta\mathbf{p}_c - [R_{wc}\mathbf{p}_c]_\times \delta\boldsymbol{\theta} + \delta\mathbf{t}$$

Depth Z	Inverse-depth error δρ	First-order |δZ|
1 m	0.01 m⁻¹	10 mm
2 m	0.01 m⁻¹	40 mm
3 m	0.01 m⁻¹	90 mm

> A fixed inverse-depth error produces a depth error proportional to Z squared. Back-projection adds pixel and camera-calibration uncertainty. At one metre, a one-degree rotation error already contributes about 17.5 mm laterally. Unknown space is not free space. An SDF built from uncertain monocular predictions needs explicit validity and uncertainty handling before it can support a control clearance decision.

量纲核验 / Units：逆深度以 m⁻¹、Z 以 m、角度微扰以 rad。17.5 mm 来自 1 m × sin(1°)，不是论文实测。

10 / 对接 S9：表面配准比动作生成器更关键

本机 V:\embodied-3d-affordance-policy\results\README.md 与 DELIVERY_REPORT；S9 日志

S9 在 64 个配对场景中引入 45° 相机视角变化：带训练拟合表面观察器的世界系点云 Flow 成功 59/64（92.2%），RGB-D 图像 Flow 为 0/64。这个结果不能被简写成“点云策略天然免标定”——高成功率主要来自已标定点云和表面配准；同条件 DDIM 与 Flow 成功数相同。移除观察器后，神经策略独立承担几何识别，成绩明显改变。

如果以 DA V2 替换原 RGB-D 的 D，变化不只是一张输入通道：相对尺度、非均匀深度误差和细结构置信度都进入观察器。需冻结相同策略和场景，先比较真实 D、带噪 D、metric DA 分支以及经过稀疏锚点校准的 relative 分支。逐项记录表面配准残差、定位误差、终端姿态和碰撞，而不是只检查网络 forward 能否接通。

新增障碍实验也有清晰机制：同一检查点开启完整 SDF 引导后 64/64 成功、0/64 碰撞；清除碰撞主要靠 Euler 步内对干净路径的近端修正，不是仅靠流梯度。该结果依赖本机合成场景里的几何观察与障碍定义，不能直接迁移成单目预测 SDF 的安全保证。建议 DA 先作为几何补全先验，metric 传感器和已测标定继续承担尺度锚。

S9 evidence	数字	解读
45° registered point Flow	59/64 success	表示与观察器联合效果
45° RGB-D image Flow	0/64 success	图像外参脆弱
Added obstacle + full guidance	64/64 success; 0 collision	固定检查点测试期修正
Latency without / with guidance	40.37 / 184.36 ms	五步、单样本中位数

```
RGB -> DA raw inverse score -> affine/metric calibration
    -> validity mask -> camera back-projection
    -> world transform -> surface registration
    -> frozen S9 policy -> swept-tool clearance checks
Report: registration residual, pos_mm, rot_deg, collision, success
```

> S9's 59/64 success under a 45-degree camera shift includes calibrated point clouds and surface registration. Flow alone does not explain the gain. Replacing measured depth with DA V2 changes the observer contract: metric scale, spatial bias and confidence must be retested. The successful SDF intervention relies on proximal path corrections and known scene geometry; it is not evidence that an arbitrary monocular SDF is collision-safe.

11 / 对接 S3、S7、S8：把观察误差和控制误差分开

本机 S3/S7/S8 报告；本册跨模块实验设计

S3 的非抓持推块实验证明：单步动作头足够快，仍可能在开环专家数据之外丢失接触。第 32 tick 首动作误差从专家观测约 17.8 mm 扩大到自身闭环约 234.3 mm。若用 DA 深度增强观测，应保留相同恢复状态集，比较接触几何估计是否改善，而不是把“多了深度”与“学会恢复”画等号。相对坐标和可达恢复状态覆盖仍是独立的控制问题。

S7 的生成 SWD 衡量光学图像特征分布，不能拿来当深度排序准确率，也不能由编译加速推算 DA V2 的相同倍数收益。真正部署时需用 DA 所选编码器、输入分辨率和后处理路径实测，包括内存、上采样与主机拷贝。原论文 V100 的速度数字只描述原作环境，不能写成 RTX 4060 Ti 的结果。

S8 的事件流可补强高速运动线索，却未天然提供单目 metric 深度。与 DA 融合前，应先将事件时钟与 RGB 曝光、相机内参和原图网格统一；在动态区域，瞬时深度与曝光平均 RGB 的不一致会给稠密几何带来结构误差。建议先做静态标定场景，再逐步增加速度，最后检验涡量、接触与安全等目标任务，而不是一次混合后只报告一个总分。

模块	保持不变	只改变并测量
S3 接触恢复	同一策略/恢复状态/场景	深度来源 → 接触定位和成功率
S7 编译比较	同权重/分辨率/精度	执行器 → 时延/数值差
S8 高速感知	同轨迹/同曝光/同标定	深度/事件融合 → 几何误差
S9 外参鲁棒	同世界场景/同策略	D 来源 → 配准/碰撞/成功

> Use a paired experiment to separate perception changes from policy changes. S3 still requires recovery-state coverage; S7's SWD and compiler speedups do not transfer to a depth model; S8's event timing does not supply metric scale by itself. Freeze downstream components first, verify coordinate and timestamp contracts, then measure the relevant task-specific quantities under controlled interventions.

12 / 许可、术语与最小使用清单

论文 arXiv 许可；官方 README 的 LICENSE；本册 CPU 核验

许可证按对象分别记录。论文 arXiv 页面是非独占分发许可；官方代码仓库采用 Apache-2.0，README 单独说明 Small 权重为 Apache-2.0，Base/Large/Giant 权重为 CC BY-NC 4.0。权重许可不授予论文全文翻译权，论文许可也不自动覆盖权重商业用途。本包不下载权重，不复制整篇中文替代文本；原创中英推导与必要原图分析保留明确来源。

math_checks.py 检验正仿射变换下的归一化不变性、两参数逆深度校准解及 Z^2 误差放大。CPU 检验结果、PDF 页数和抽检图在 verification.json 与 work/ 中。网络张量形状同时对照原论文和官方 dpt.py；没有运行模型推理，不能把静态形状核对写成模型部署测试通过。

最小使用清单：确认使用 relative 还是 metric checkpoint；保存原始浮点图；保留 resize/padding 到相机网格的变换；在实际距离范围检验尺度和误差；世界系融合使用已测外参；把无效与遮挡空间标成未知；最后在冻结控制器上做配对闭环验收。阅读路线为本册 1→4→6→7→9→10，再读原文附录 B.7/B.9 与 C 的评测选样细节。

English term	本册固定用法
affine-invariant inverse depth	仿射不变逆深度
metric depth	米制深度
gradient matching	梯度匹配
pseudo-labeled real images	伪标注真实图像
surface registration	表面配准

> Licenses attach to different objects: the paper, repository code and individual weight families. Do not transfer permissions between them. The included CPU checks establish algebraic invariants and calibration calculations, not deployment quality. Before downstream use, choose the correct checkpoint family, preserve raw values and image transforms, validate metric scale, and test the unchanged controller on paired task scenes.

References / 阅读来源：Yang et al. Depth Anything V2. NeurIPS 2024. arXiv:2406.09414v2. DOI:10.48550/arXiv.2406.09414. Official implementation: github.com/DepthAnything/Depth-Anything-V2. 另见本机 embodied-3d-affordance-policy、event-neural-flow、rectified-flow-distill-bench 与 S3 闭环实验报告。