7 Training Methodology

Teacher-Student Paradigm

Teacher: DINOv2+DPT

- **Monocular only** (single images)
- Standard DPT decoder
- No multi-view / No camera tokens
- Trained on synthetic data only

Student: DA3

- **Multi-view** (2-18 images)
- Dual-DPT + cross-view attention
- Camera tokens + camera head
- Trained on mixed real+synthetic

Synthetic → Teacher → Relative Depth

RANSAC align (s,t) → cached pseudo-labels

↓

Real+Noisy → Pseudo-Labels → Student (DA3)

Mixed Data Strategy

Real Depth | 3D Recon | Synthetic

~40% synthetic, ~30% real (ScanNet++), ~30% reconstructed (DL3DV)

KEY DETAILS

200k steps on 128 H100 GPUs

- Base res 504^2 (varied AR)
- Pose cond prob p=0.2
- Wall-clock: ~5-7 days
- Camera head D_C: ~1ms

Student Training Objective

$$L = L_D + L_M + L_P + \beta L_C + \alpha L_{grad}$$

(where $\alpha = 1, \beta = 1$)

Student: Multi-view depth + rays + pose (mixed real+synthetic data)

Teacher: $L_T = L_{grad} + L_{gl} + 0.5 L_N + L_{sky} + L_{obj}$ (monocular, synthetic only)

Student Loss Components

1. Confidence-Aware Depth Loss (L_D)

Weighted L1 loss with learned uncertainty $D_{c,p}$ that down-weights ambiguous regions.

2. Ray Map Loss (L_M)

L1 loss on ray vectors $M = (t, d)$ for direction consistency.

3. Point Consistency (L_P)

Loss on 3D points $P = t + \hat{D} \odot d$ to enforce geometric validity.

4. Camera Head (L_C)

Supervision for pose predictions (f, q, t) from camera head.

5. Gradient Loss (L_{grad})

Edge-aware smoothness (shared with teacher).

Teacher Losses (monocular)

ROE, surface normals, sky/object masks