π^3: Permutation-Equivariant Visual Geometry Learning

Yifan Wang[1*] Jianjun Zhou[123*] Haoyi Zhu[1] Wenzheng Chang[1] Yang Zhou[1]
Zizun Li[1] Junyi Chen[1] Jiangmiao Pang[1] Chunhua Shen[2] Tong He[13†]
[1]Shanghai AI Lab [2]ZJU [3]SII
*Equal Contribution †Corresponding Author

- Problem / objective
  - Task: Visual Geometry Reconstruction
  - Problem: Previous research's over-reliance on fixed reference view, where its inductive bias leads to instability and failures if the reference is suboptimal.
- Contribution / Key idea
  - π3: Fully permutation-equivariant architecture that eliminates reference-view based bias
    i. Predict affine-invariant camera poses
    ii. Predict scale-invariant pointmaps

- **Overview**

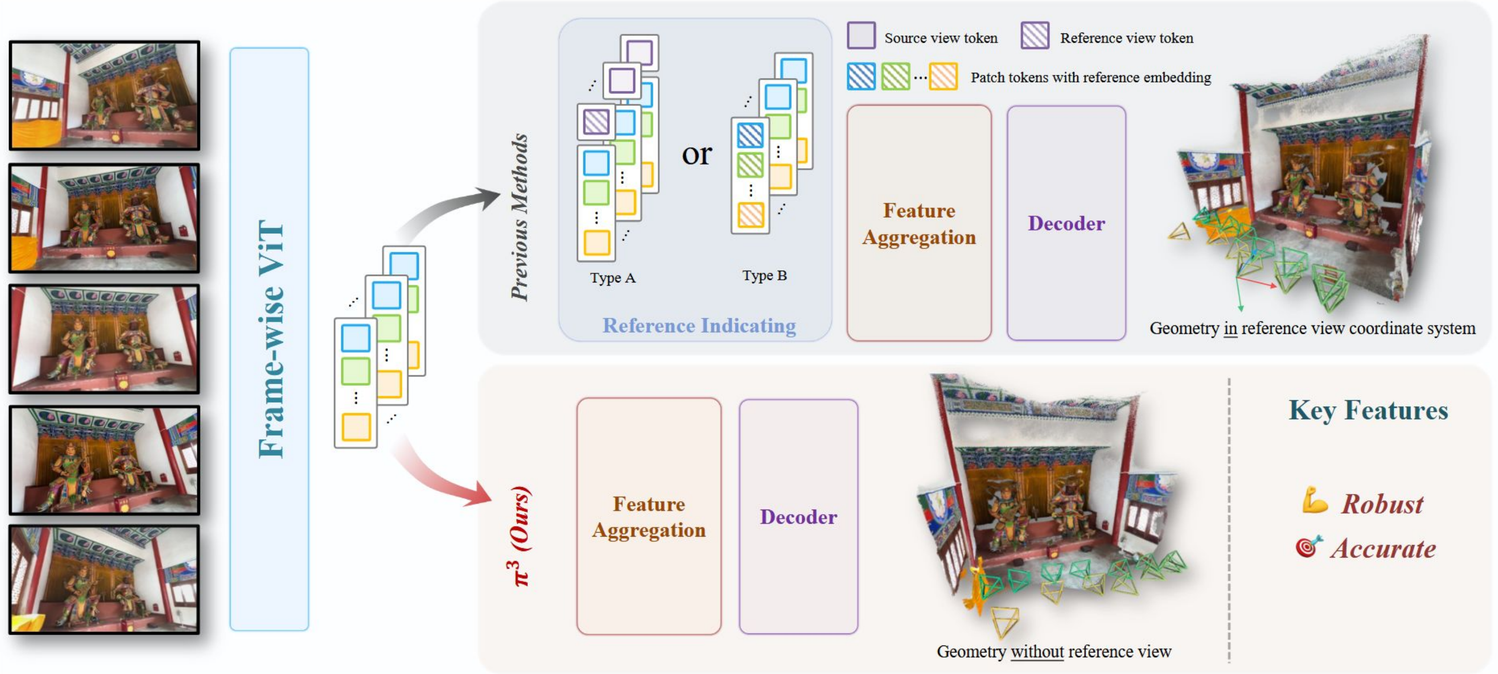

Figure 3. Unlike prior methods that designate a *reference view* by concatenating a special token (Type A) or adding a learnable embedding (Type B), π^3 achieves permutation equivariance by eliminating this requirement altogether. Instead, it employs relative supervision, making our approach inherently robust to the order of input views.

- **Permutation-Equivariant Architecture**

1. Model Input $S = (\mathbf{I}_1, \ldots, \mathbf{I}_N)$
2. Model Output $\phi(S) = ((\mathbf{T}_1, \ldots, \mathbf{T}_N), (\mathbf{X}_1, \ldots, \mathbf{X}_N), (\mathbf{C}_1, \ldots, \mathbf{C}_N))$
3. Permutation-Equivariance

$$\phi(P_\pi(S)) = P_\pi(\phi(S)) \tag{2}$$

$$\begin{aligned} P_\pi(\phi(S)) = \big(&(\mathbf{T}_{\pi(1)}, \ldots, \mathbf{T}_{\pi(N)}), \\ &(\mathbf{X}_{\pi(1)}, \ldots, \mathbf{X}_{\pi(N)}), \\ &(\mathbf{C}_{\pi(1)}, \ldots, \mathbf{C}_{\pi(N)})\big) \end{aligned} \tag{3}$$

- **Scale-Invariant Local Geometry**

1. Point cloud reconstruction loss
    a. Problem: Inherent scale ambiguity challenge in monocular reconstruction
    b. Solution: Find a single optimal scale factor minimizing depth-weighted L1 distance loss

$$s^* = \arg\min_s \sum_{i=1}^{N} \sum_{j=1}^{H \times W} \frac{1}{z_{i,j}} \|s\hat{\mathbf{x}}_{i,j} - \mathbf{x}_{i,j}\|_1 \qquad (4)$$

$$\mathcal{L}_{\text{points}} = \frac{1}{3NHW} \sum_{i=1}^{N} \sum_{j=1}^{H \times W} \frac{1}{z_{i,j}} \|s^*\hat{\mathbf{x}}_{i,j} - \mathbf{x}_{i,j}\|_1 \qquad (5)$$

2. Normal loss
    a. Objective: Reconstruction of locally smooth surface

$$\mathcal{L}_{\text{normal}} = \sum_{i=1}^{N} \sum_{j=1}^{H \times W} \arccos(\hat{\mathbf{n}}_{i,j} \cdot \mathbf{n}_{i,j}) \qquad (6)$$

- **Affine-Invariant Camera Pose**

1. Camera loss = Rotation Loss + Translation loss

$$\mathcal{L}_{\text{cam}} = \frac{1}{N(N-1)} \sum_{i \neq j} \left(\mathcal{L}_{\text{rot}}(i,j) + \lambda_{trans} \mathcal{L}_{\text{trans}}(i,j) \right) \tag{8}$$

$$\mathcal{L}_{\text{rot}}(i,j) = \arccos \left(\frac{\text{Tr}\left((\mathbf{R}_{i \leftarrow j})^{\top} \hat{\mathbf{R}}_{i \leftarrow j} \right) - 1}{2} \right) \tag{9}$$

$$\mathcal{L}_{\text{trans}}(i,j) = \mathcal{H}_{\delta}(s^{*} \hat{\mathbf{t}}_{i \leftarrow j} - \mathbf{t}_{i \leftarrow j}) \tag{10}$$

- **Model Training**

  1. Final loss = Point reconstruction loss + Confidence loss + Camera pose loss

$$\mathcal{L} = \mathcal{L}_{\text{points}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}} + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}} + \lambda_{\text{cam}}\mathcal{L}_{\text{cam}} \quad (11)$$

- **Experiments**

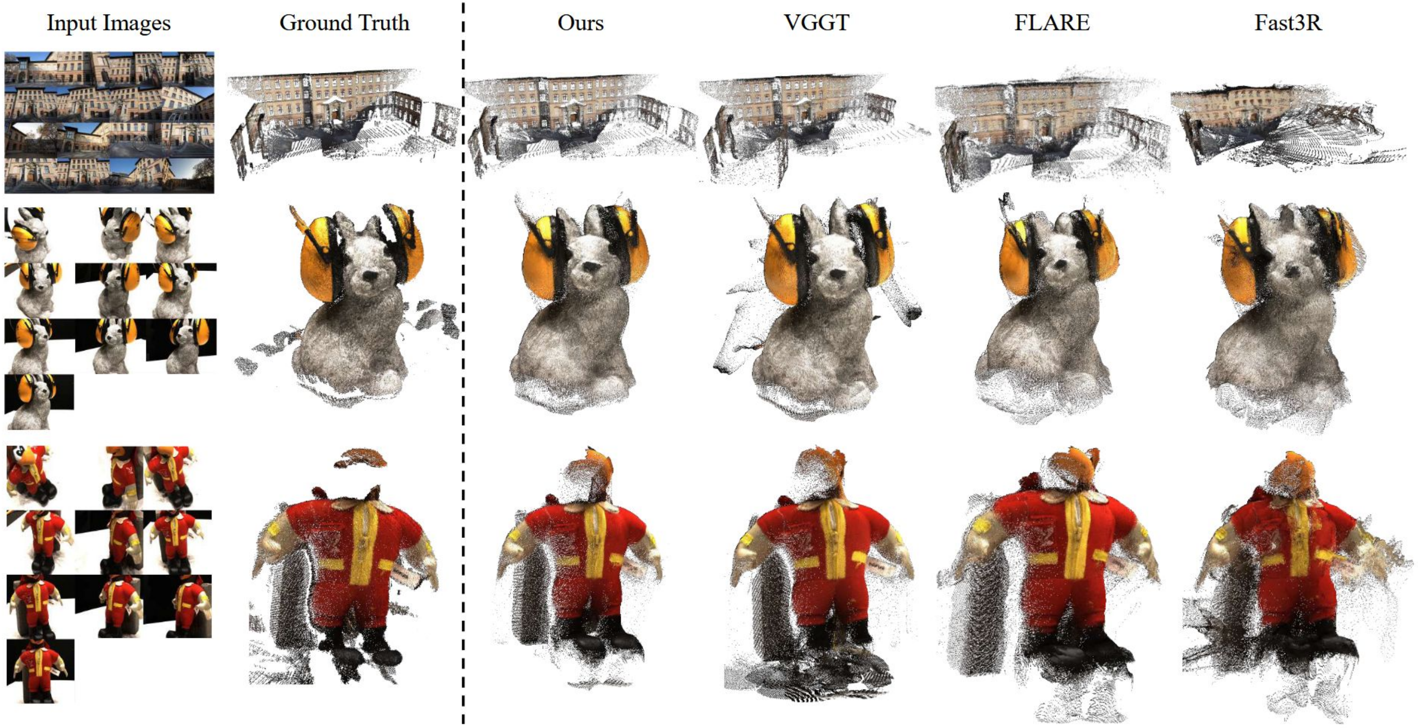

Figure 5. **Qualitative comparison of multi-view 3D reconstruction.** Compared to other multi-frame feed-forward reconstruction methods, π^3 produces cleaner, more accurate and more complete reconstructions with fewer artifacts.