

Problem Setup



Arbitrary Visual Inputs

Input set could be given as a single image, stereo pair, or video frame sequence.



Per-Pixel Ray Map M

Instead of global pose, the model predicts a dense map M containing ray origin t and direction d for every pixel.

Why Avoid Explicit Rotation?

Directly regressing a rotation matrix R imposes an **orthogonality constraint** ($R^T R = I$) that is difficult for neural networks to satisfy, often leading to optimization instability or degenerate poses.

- ✓ Ray directions $d \in \mathbb{R}^3$ avoid explicit orthogonality constraints, simplifying optimization.

Key: Ray direction is derived from classical formulation as $d = RK^{-1}p$, linking rotation matrix and intrinsics to rays.

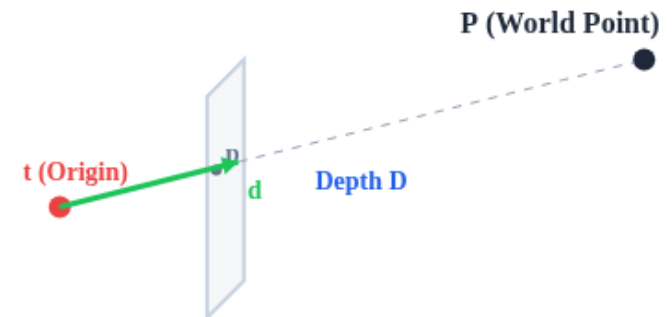
Why is d unnormalized?

d is unnormalized because $\|d\|$ encodes the projection transformation from K^{-1} . Normalizing would decouple the scale relationship between depth and projection. The unnormalized d preserves the pixel-to-world scale factor, which is critical for geometric consistency: $P = t + D \cdot d$ requires d 's magnitude to properly fuse depth with ray direction.

$$P = t + D(u, v) d$$

$$P \text{ (3D Point)} = t \text{ (Ray Origin)} + D \text{ (Depth)} \times d \text{ (Direction)}$$

GEOMETRIC DERIVATION



The direction d derived via backprojection: $d = RK^{-1}p$