Image Input

Ground Image
I_g

Aerial Image
I_a

Token Extraction
$\mathbf{F}_g \in \mathbb{R}^{N \times D}$

Token Extraction
$\mathbf{F}_a \in \mathbb{R}^{N \times D}$

DINOv2/CLIP/ResNet50

Backbone Agnostic

Sky Filtering & Depth
$M_{grid}, d_{i,j}$

Dynamic Grid Size
$G = \sqrt{N}$

Core CroDINO Method

3 Distance Layers

3 Depth Layers

Vertical Aggregation
Ground Image
$\mathbf{t}_j^{fore/mid/back}$

Depth-weighted:
$w_r^{fore} = 1 - r/R$
$w_r^{back} = r/R$

Radial Aggregation
Aerial Image
$\mathbf{r}_\beta^{fore/mid/back}$

Cross-Modal Alignment
Cosine Distance Minimization

$\mathcal{L}(\theta) = \frac{1}{G} \sum_{i=0}^{G} \left\| 1 - \mathbf{t}_i^T \mathbf{r}_{\phi(\theta, i)} \right\|$

Orientation Estimate
$\theta^* = \arg\min_\theta \mathcal{L}(\theta)$