LoDEOT: Low-Dimensional and Efficient Offset Tokens for Building Footprint Extraction from Off-Nadir Imagery

Anonymous authors
Paper under double-blind review

Abstract

Instance-level roof-to-footprint offset (RFO) prediction is central to extracting building footprints from off-nadir imagery. Query-based pipelines commonly use high-dimensional instance tokens to predict signed two-dimensional RFOs. We investigate whether RFO prediction can instead use a compact offset token. Under local pinhole projection and vertical-extrusion assumptions, the idealized RFO map admits a five-parameter sufficient descriptor comprising intrinsic shape, composite amplitude, and relative geometry. This factorization provides a structural prior for a five-dimensional offset token, whose channels learn task-relevant latent representations through end-to-end training. Based on this design, we propose LoDEOT, which retains high-dimensional instance tokens for detection and segmentation but maps instance-token, concentration-gated roof, and box–mask evidence to a five-dimensional offset token followed by an independent two-dimensional readout. Known denoising-query target indices further align each supervised decoder-layer estimate with the same clean instance RFO, organizing successive predictions as target-aligned recovery under perturbed query conditions. Experiments on five real-world building datasets demonstrate the effectiveness of LoDEOT for building footprint extraction. Experiments on real-world building datasets demonstrate that a five-dimensional offset token can support accurate RFO prediction. On BONAI, LoDEOT achieves the best roof-detection bAP and bAP50 and leads all five offset-corrected footprint metrics among the evaluated end-to-end methods, with FAP50 of 54.58 and mEPE of 5.23 pixels. Its FAP50 exceeds those of the evaluated end-to-end baselines by 7.56–16.85 percentage points.

1 Introduction

Accurate building footprints support urban mapping, population analysis, disaster response, and 3D city modeling. Raster segmentation and polygonal extraction improve building delineation (Maggiori et al., 2017; Demir et al., 2018; Ji et al., 2019; Li et al., 2021; Girard et al., 2021), but viewing direction and building height displace visible roofs from their ground outlines in off-nadir imagery. The signed roof-to-footprint offset (RFO) describes this roof-to-footprint displacement in the image plane; its accurate prediction supports footprint localization, boundary alignment, height estimation, and 3D building reconstruction (Wang et al., 2023; Li et al., 2024b).

RFO prediction has progressed from convolutional roof–offset branches, including LOFT and MLS-BRN (Wang et al., 2023; Li et al., 2024b), to prompt-driven and token-based instance models (Li et al., 2024a; 2025a). DG-BRF combines diffusion-based roof and facade segmentation with geometric priors for offset estimation (Hu et al., 2026). These methods improve observations, supervision, or output structure, yet the offset representation used to predict the offset itself remains largely unexamined, *e.g.,* commonly used 256, 512, 1024 dimensions for token design (Li et al., 2024a; 2025a;b; 2026b). A generic instance token must stay expressive for classification, localization, and segmentation; forcing the task-specific offset route to inherit the same width adds width-dependent offset-head parameters, while roof appearance, mask reliability, and box–mask asymmetry remain only implicitly available. The issue is not whether to compress the shared token, but whether the RFO route can be decoupled into a compact offset head.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Three observations guide our design. First, prior work shows that Transformer representations can encode geometric relationships (Rombach et al., 2021), camera pose (Bahmani et al., 2025), and compressible dense geometric information (Zhou et al., 2026). These findings motivate investigating geometry-informed RFO tokens. Second, under local pinhole-camera, vertical-extrusion, and visibility assumptions, principal-point cancellation, coordinate gauges, scale coupling, and consolidation of relative geometric factors yield an exact five-parameter sufficient description of the idealized RFO map. This factorization motivates a five-dimensional candidate for the offset head; it does not establish a unique, minimal, or universally optimal neural width. Third, successive offset decoder layers support target-aligned recovery of the offset tokens: the offset tokens are gradually close to the ground truth offset. This is a layer-by-layer denoising process.

Guided by these observations and a first-principles analysis of the mathematical relationship between RFO and physical parameters, i.e., building height & camera parameters, we propose an RFO prediction framework centered on low-dimensional and efficient offset tokens (LoDEOT), a five-dimensional offset token representation. As the forward process of offset token exhibit a natural denoising feature, we design a strategy to instill random noise to enhance the training.

On BONAI (Wang et al., 2023), LoDEOT achieves the highest roof-detection bAP and bAP50 among the compared end-to-end methods and leads all five offset-corrected footprint metrics, reaching 54.58% FAP50 and a 5.23-pixel mEPE. Statistical analysis finds no significant performance differences in the tested width comparisons; we retain the geometry-motivated five-dimensional representation as a compact choice. Target-indexed denoising improves all reported ablation metrics, while the largest cumulative source set achieves the best five-domain macro-average results.

The contributions of this work are threefold:

- Starting from first principles, we analyze the intrinsic relationship between the RFO and key physical parameters in the photogrammetric process, and tailor this process for off-nadir remote sensing scenario.
- Combining the photogrammetric analysis with domain insights into modern foundation models, we propose LoDEOT and equip it with a customized denoising training framework.
- Based on our design, we conduct extensive empirical evaluations to demonstrate the practical effectiveness of our method, and further corroborate our initial observations through the experimental phenomena.

2 Related Work

Raster and Polygonal Building Extraction. Building extraction has progressed from semantic raster prediction on cross-city and large-scale benchmarks (Maggiori et al., 2017; Demir et al., 2018; Ji et al., 2019) to explicit geometric outputs. Polygon-RNN++ recurrently predicts vertices and refines them with graph reasoning (Acuna et al., 2018); joint semantic–geometric learning combines masks, corners, and edge orientations (Li et al., 2021). Frame Field Learning predicts contour directions for polygonization (Girard et al., 2021), whereas PolyWorld learns vertex connectivity through graph reasoning and differentiable matching (Zorzi et al., 2022). CVNet evolves building contours (Xu et al., 2022); BuildMapper learns initialization, deformation, and vertex reduction (Wei et al., 2023); and HiSup couples vertex, boundary, and mask supervision (Xu et al., 2023). PolyBuilding uses polygon queries and fixed-length point sequences (Hu et al., 2023), while Pix2Poly generates polygon vertices and connectivity as a sequence (Adimoolam et al., 2025). These methods improve boundary geometry but do not directly resolve the roof–footprint displacement specific to off-nadir imagery.

Off-Nadir Roof-to-Footprint Learning. LOFT jointly predicts roof masks and RFOs and introduced BONAI (Wang et al., 2023); a multi-label ViT jointly models footprints, roofs, and visible building shapes (Neupane et al., 2025). OBM makes roof/offset extraction prompt driven (Li et al., 2024a), MLS-BRN uses multi-level supervision and a roof-offset-guided footprint extractor (Li et al., 2024b), and PolyFootNet combines direct polygon prediction with mask prompting and self-offset attention (Li et al., 2025a). Building on the promptable Segment Anything Model (Kirillov et al., 2023), RBF-SAM improves robustness to prompt type and quality (Li et al., 2026b). These advances change observations, outputs, or supervision. LoDEOT instead studies the capacity

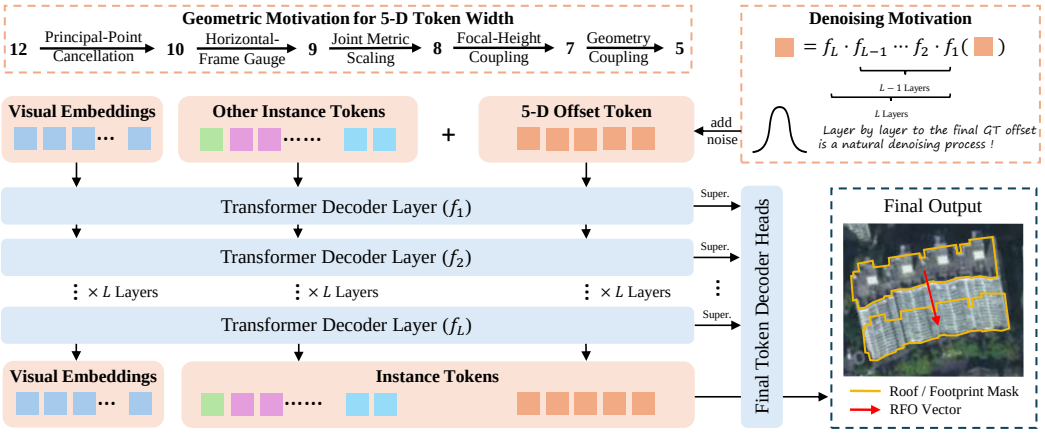


Figure 1: RFO prediction architecture after incorporating LoDEOT. Motivated by the projective geometry of 3D imaging, we redesign the offset-token representation. After the backbone encodes the visual features, each original high-dimensional instance token is augmented with a 5-D LoDEOT that carries the task-specific RFO representation, while retaining its roles in classification, roof-mask, and roof-box prediction. The augmented instance tokens are progressively updated through L Transformer decoder layers; the final head predicts roof masks and RFOs to recover footprints. During training, each layer's RFO prediction is supervised by the same clean instance target.

of the offset head itself. LOFT and MLS-BRN are the directly comparable roof-mask/RFO baselines under the reported BONAI protocol; prompt-dependent and polygon-output methods use different interfaces or evaluation settings.

Query-Based Representation and Denoising. Instance segmentation progressed from Mask R-CNN's region-conditioned mask branch (He et al., 2017) to QueryInst's shared queries for categories, boxes, masks, and associations (Fang et al., 2021). DETR formulates detection as set prediction with object queries and bipartite matching (Carion et al., 2020); Conditional, Deformable, and DAB-DETR spatially condition them through conditional queries, sparse reference-point sampling, and dynamic anchors (Meng et al., 2021; Zhu et al., 2021; Liu et al., 2022). MaskFormer uses mask classification, and Mask2Former restricts attention to predicted mask regions (Cheng et al., 2021; 2022). DN-DETR reconstructs perturbed labels and boxes with known correspondence (Li et al., 2022); DINO adds contrastive denoising (Zhang et al., 2023), and Mask DINO unifies detection and segmentation with mask-enhanced denoising (Li et al., 2023). Existing denoising objectives focus on classes, boxes, and masks. LoDEOT updates a compact offset token at each layer by combining the previous offset token with current instance evidence, and uses known denoising target correspondences to supervise each layer's prediction against the same clean instance RFO.

3 Method

LoDEOT predicts roof instances and signed RFOs using high-dimensional instance tokens (decoder queries) and associated compact offset tokens (Figure 1).

3.1 Overall Architecture

Given an off-nadir image $\mathbf{I}$, an image encoder and an L-layer instance decoder produce K query-indexed predictions:

$$\widehat{\mathcal{Y}} = \left\{ \left(\hat{c}_q, \hat{\mathbf{b}}_q, \hat{\mathbf{m}}_q, \hat{\boldsymbol{\delta}}_q \right) \right\}_{q=1}^{K}. \tag{1}$$

Here $\hat{c}_q$, $\hat{\mathbf{b}}_q$, and $\hat{\mathbf{m}}_q$ denote the class, roof box, and roof mask. The signed RFO $\boldsymbol{\delta} = \mathbf{u}_b - \mathbf{u}_r \in \mathbb{R}^2$ points from a roof reference point $\mathbf{u}_r$ to its corresponding footprint reference point $\mathbf{u}_b$. Both points and the displacement are expressed in the same pixel coordinate system.

Footprints are recovered by translating each predicted roof mask by $\hat{\boldsymbol{\delta}}_q$ under a local constant-displacement approximation. Multitask instance tokens retain their width; only offset-token width d varies. Each offset token is updated from its previous value and current instance evidence.

3.2 Geometry-Inspired Candidate Width

RFO depends on camera intrinsics, relative camera–building pose, and building height through shared combinations. We derive an exact five-parameter factorization of this map to guide offset-token design.

Explicit Physical Offset Computation. Within a local instance neighborhood, let $\mathbf{B} \in \mathbb{R}^3$ be a footprint reference point, $\mathbf{C} \in \mathbb{R}^3$ the camera center, $\mathbf{g} \in \mathbb{S}^2$ the fixed world-vertical direction, and $H > 0$ the building height. Vertical extrusion gives the roof point $\mathbf{B} + H\mathbf{g}$, while $\mathbf{r} = \mathbf{B} - \mathbf{C}$ removes the arbitrary world origin. Let $\mathbf{R} \in \mathrm{SO}(3)$ map world directions to the camera frame. A pinhole intrinsic matrix $\mathbf{K}$, normalized by $K_{33} = 1$, contains focal scales $f_x, f_y > 0$, principal point (c_x, c_y), and image-axis skew κ. With $\pi([a, b, c]^\top) = [a/c, b/c]^\top$, the two projections and their signed difference are

$$\begin{aligned} \mathbf{u}_b &= \pi(\mathbf{KRr}), \\ \mathbf{u}_r &= \pi(\mathbf{KR}(\mathbf{r} + H\mathbf{g})), \\ \boldsymbol{\delta} &= \mathbf{u}_b - \mathbf{u}_r = \Phi(\mathbf{K}, \mathbf{R}, \mathbf{r}, H). \end{aligned} \tag{2}$$

The local state has twelve coordinates: five intrinsics, three rotation coordinates, three components of $\mathbf{r}$, and H.

Five-parameter factorization of the physical map. Define camera-frame vectors $\boldsymbol{\xi} = \mathbf{Rr} = (X, Y, Z)^\top$ and $\boldsymbol{\nu} = \mathbf{Rg} = (\nu_x, \nu_y, \nu_z)^\top \in \mathbb{S}^2$, normalized coordinates $x = X/Z$, $y = Y/Z$, and relative height $\rho = H/Z$. For both reference points to lie in front of the camera, require $Z > 0$ and $Z + H\nu_z > 0$. Subtracting their projections yields

$$\delta_x = \frac{\rho}{1 + \rho\nu_z}[f_x(x\nu_z - \nu_x) + \kappa(y\nu_z - \nu_y)], \tag{3}$$

$$\delta_y = \frac{f_y\rho}{1 + \rho\nu_z}(y\nu_z - \nu_y). \tag{4}$$

Shared camera intrinsics make the principal point (c_x, c_y) cancel from the projection difference, reducing the coordinate count from twelve to ten.

Pose enters only through $(\boldsymbol{\xi}, \boldsymbol{\nu})$. For $\mathbf{Q} \in \mathrm{SO}(3)$ with $\mathbf{Qg} = \mathbf{g}$, the transformation $(\mathbf{R}, \mathbf{r}) \mapsto (\mathbf{RQ}^\top, \mathbf{Qr})$ preserves both vectors. This horizontal-frame gauge preserves physical viewing geometry and replaces six pose coordinates with three for $\boldsymbol{\xi}$ and two for $\boldsymbol{\nu} \in \mathbb{S}^2$, leaving nine overall.

Joint scaling $(\mathbf{r}, H) \mapsto (\lambda\mathbf{r}, \lambda H)$, $\lambda > 0$, preserves projections; (X, Y, Z, H) enters only through $(x, y, \rho) = (X/Z, Y/Z, H/Z)$, leaving eight coordinates: three intrinsics, (x, y, ρ), and two for $\boldsymbol{\nu}$.

Define focal anisotropy $a = f_x/f_y$, normalized skew $\widetilde{\kappa} = \kappa/f_y$, and the composite amplitude $\gamma = f_y\rho/(1 + \rho\nu_z)$ combining common focal scale and relative height. Then

$$\boldsymbol{\delta} = \gamma \begin{bmatrix} a & \widetilde{\kappa} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x\nu_z - \nu_x \\ y\nu_z - \nu_y \end{bmatrix}. \tag{5}$$

To make this coupling explicit, let $\mathbf{A} = \left[\begin{smallmatrix} f_x & \kappa \\ 0 & f_y \end{smallmatrix}\right]$. For any $\mu > 0$ satisfying $\mu + (\mu - 1)\rho\nu_z > 0$, set $\mathbf{A}' = \mu\mathbf{A}$ and $\rho' = \rho/[\mu + (\mu - 1)\rho\nu_z]$. This transformation preserves positive depth and satisfies

$$\frac{\rho'}{1 + \rho'\nu_z}\mathbf{A}' = \frac{\rho}{1 + \rho\nu_z}\mathbf{A}. \tag{6}$$

Thus, at fixed $(x, y, \boldsymbol{\nu})$, common focal scale and relative height affect the offset only through γ. The retained variables are $(a, \widetilde{\kappa}, \gamma, x, y, \boldsymbol{\nu})$, with seven coordinates in total.

After introducing γ, the remaining positional and directional dependence enters through the geometric vector

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_x \\ \beta_y \end{bmatrix} = \begin{bmatrix} x\nu_z - \nu_x \\ y\nu_z - \nu_y \end{bmatrix}. \tag{7}$$

The four coordinates of $(x, y, \boldsymbol{\nu})$ are therefore represented by two combinations while γ is retained separately. For fixed $\boldsymbol{\nu}$ with $\nu_z \neq 0$, the map from (x, y) to $\boldsymbol{\beta}$ has Jacobian $\nu_z \mathbf{I}_2$, so these two geometry coordinates can vary independently. Combining the preceding steps gives

$$\begin{aligned} 12 &\xrightarrow{-2 \text{ principal point}} 10 \xrightarrow{-1 \text{ horizontal gauge}} 9, \\ 9 &\xrightarrow{-1 \text{ metric scale}} 8 \xrightarrow{-1 \text{ focal–height}} 7 \xrightarrow{-2 \text{ geometry}} 5, \\ \boldsymbol{\eta}_5 &= (a, \widetilde{\kappa}, \gamma, \beta_x, \beta_y)^\top, \\ \boldsymbol{\delta} &= \Phi_5(\boldsymbol{\eta}_5) = \gamma \begin{bmatrix} a & \widetilde{\kappa} \\ 0 & 1 \end{bmatrix} \boldsymbol{\beta}. \end{aligned} \tag{8}$$

The counts describe retained factorization coordinates, not a minimum representation dimension. The description is exact under the pinhole-camera, vertical-extrusion, and positive-depth assumptions, including $\nu_z = 0$, where $\boldsymbol{\beta} = -(\nu_x, \nu_y)^\top$ has unit norm and geometry is more restricted.

From Physical Factorization to the Default Offset-Token Width. The five-parameter descriptor $\boldsymbol{\eta}_5$ organizes the general imaging map into intrinsic shape, displacement amplitude, and relative geometry, motivating $d = 5$ as the default offset-token width. Learned channels need not correspond individually to physical parameters, and the readout need not implement Φ_5. Under shared calibration and locally restricted viewing geometry, some factors can become shared or coupled, providing a possible explanation for similar performance at smaller widths. The supplementary material derives these conditional reductions, which inform the width analysis in Section 4.4.

3.3 Mask-Guided Compact Offset Head

We omit the query index and show layer subscripts only in recurrent updates; all vectors are columns. On a resized grid of at most $S \times S$ with P locations, $\mathbf{F}, \mathbf{E} \in \mathbb{R}^{P \times d_m}$ denote mask features and sinusoidal positional encodings, and $\mathbf{m} \in \mathbb{R}^P$ roof-mask logits. With sigmoid σ and stop-gradient operator sg, set $\mathbf{w} = \sigma(\mathrm{sg}(\mathbf{m}))$ and pool features as

$$\mathbf{p} = \frac{\mathbf{w}}{\mathbf{1}^\top \mathbf{w} + \epsilon}, \qquad \mathbf{f} = f_{\text{feat}}\big((\mathbf{F} + \mathbf{E})^\top \mathbf{p}\big). \tag{9}$$

Here $\mathbf{1} \in \mathbb{R}^P$ is the all-ones vector, $\epsilon > 0$ stabilizes normalization, and the feature branch remains trainable. During training, selected queries receive zero-mean Gaussian noise on $\mathbf{w}$, clipped to $[0, 1]$ before normalization.

Let $\mathbf{X} \in [0, 1]^{P \times 2}$ contain normalized grid coordinates, and let $\mathbf{b} \in \mathbb{R}^4$ be the detached normalized center–size roof box with center $\mathbf{b}_c \in \mathbb{R}^2$. The stabilized mask centroid is $\mathbf{c} = \mathbf{X}^\top \mathbf{p}$. Spatial evidence is constructed as

$$s = \|\mathbf{p}\|_\infty \exp\left(\frac{\mathbf{p}^\top \log(\mathbf{p} + \epsilon \mathbf{1})}{P}\right), \qquad \boldsymbol{\zeta} = f_{\text{sp}}\big(\mathrm{Concat}\big(\mathbf{b}, \mathbf{c}, s(\mathbf{c} - \mathbf{b}_c)\big)\big). \tag{10}$$

The logarithm acts elementwise, and Concat concatenates column vectors. The concentration score s gates roof features and box–centroid displacement. It is sensitive to spatial support size and is not a calibrated mask-quality estimate.

At decoder layer ℓ, OffsetHead updates the offset token using its previous value and current instance evidence:

$$\mathbf{o}_\ell = f_{\text{tok}}\big(\mathrm{Concat}\big(\mathbf{o}_{\ell-1}, \mathbf{h}_\ell, s_\ell \mathbf{f}_\ell, \boldsymbol{\zeta}_\ell\big)\big), \qquad \hat{\mathbf{z}}_\ell = f_{\text{out}}(\mathbf{o}_\ell). \tag{11}$$

Here $\mathbf{h}_\ell \in \mathbb{R}^D$ is the instance token, $\mathbf{o}_\ell \in \mathbb{R}^d$ the offset token, and $\hat{\mathbf{z}}_\ell \in \mathbb{R}^2$ the normalized RFO prediction defined below. The maps f_{feat}, f_{sp}, and f_{tok} are learned, and $f_{\text{out}} : \mathbb{R}^d \to \mathbb{R}^2$ is a linear readout. The offset token passes to the next decoder layer, with $d = 5$ as the default width.

3.4 Target-Indexed Layer-Wise RFO Supervision

Under invertible affine augmentation $T(\mathbf{u}) = \mathbf{J}\mathbf{u} + \mathbf{t}$, encode the pixel-space RFO as $\mathbf{z} = \alpha \mathbf{D}_a^{-1} \mathbf{J} \boldsymbol{\delta}$, where $\mathbf{D}_a = \mathrm{diag}(W_a, H_a)$ contains augmented image dimensions and $\alpha > 0$. For mask translation, decode $\alpha^{-1} \mathbf{D}_a \hat{\mathbf{z}}$ in augmented coordinates, applying $\mathbf{J}^{-1}$ when returning to original coordinates.

Denoising groups replicate ground-truth instances with perturbed labels and boxes. Attention masks isolate groups and block matching queries' access to denoising queries. Known target indices supervise successive RFOs against the same clean target without additional matching. For valid $\mathbf{z}_n$, a fixed lift $\mathbf{U} \in \mathbb{R}^{d\times 2}$ ($d \geq 2$, $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_2$) initializes offset tokens as $\mathbf{U}\mathbf{z}_n + \boldsymbol{\varepsilon}_q$ (additive) or $(\mathbf{U}\mathbf{z}_n) \odot (\mathbf{1}_d + \boldsymbol{\varepsilon}_q)$ (multiplicative). Here $\boldsymbol{\varepsilon}_q \in \mathbb{R}^d$ is the sampled perturbation, $\mathbf{1}_d$ the all-ones vector, and $\odot$ elementwise multiplication. Matching-query offset tokens are initialized from instance features; both query types follow Eq. (11).

For supervision, write $\hat{\mathbf{z}}_\ell(q)$ for the normalized RFO predicted by query q at layer ℓ, and $\mathbf{z}_n$ for target n. The coordinate-summed regression loss is

$$\ell_\delta(\hat{\mathbf{z}}, \mathbf{z}) = \sum_{k=1}^{2} \mathrm{SmoothL1}(\hat{z}_k - z_k), \qquad \mathrm{SmoothL1}(t) = \begin{cases} \frac{1}{2}t^2, & |t| < 1, \\ |t| - \frac{1}{2}, & |t| \geq 1. \end{cases} \tag{12}$$

Let $\mathcal{M}_\ell$ contain the Hungarian-matched query–target pairs at layer ℓ, and let $\mathcal{D}$ contain the fixed denoising pairs across all groups. With RFO-validity indicator $v_n \in \{0, 1\}$, the matching-query and denoising RFO losses are

$$\begin{aligned} \mathcal{L}_\delta(\ell) &= \frac{1}{N} \sum_{(q,n)\in\mathcal{M}_\ell} v_n \, \ell_\delta\big(\hat{\mathbf{z}}_\ell(q), \mathbf{z}_n\big), \\ \mathcal{L}_{\mathrm{DN}} &= \frac{1}{NG} \sum_{\ell\in\mathcal{S}} \omega_\ell \sum_{(q,n)\in\mathcal{D}} v_n \, \ell_\delta\big(\hat{\mathbf{z}}_\ell(q), \mathbf{z}_n\big). \end{aligned} \tag{13}$$

Here N is the ground-truth normalization count clamped to at least one, G the denoising-group count, and $\mathcal{S}$ the supervised denoising stages. Each ω_ℓ includes both stage and RFO-loss weights. Invalid RFOs ($v_n = 0$) use the placeholder $\mathbf{z}_n = \mathbf{0}$ and contribute neither regression nor matching cost, without changing N. $\mathcal{L}_{\mathrm{DN}}$ includes only RFO supervision.

Hungarian matching combines classification, point-sampled mask BCE/Dice, box L_1, and GIoU costs with the valid-only RFO cost $v_n\|\hat{\mathbf{z}}_\ell(q) - \mathbf{z}_n\|_1$, weighted by $\lambda^m_{\mathrm{box}}\lambda^m_\delta$. The joint set-prediction loss at each supervised stage is

$$\begin{aligned} \mathcal{L}_{\mathrm{set}}(\ell) = \lambda_{\mathrm{cls}}\mathcal{L}_{\mathrm{cls}}(\ell) + \lambda_{\mathrm{mask}}\mathcal{L}_{\mathrm{mask}}(\ell) + \lambda_{\mathrm{dice}}\mathcal{L}_{\mathrm{dice}}(\ell) \\ + \lambda_{\mathrm{box}}\mathcal{L}_{\mathrm{box}}(\ell) + \lambda_{\mathrm{giou}}\mathcal{L}_{\mathrm{giou}}(\ell) + \lambda_\delta\mathcal{L}_\delta(\ell). \end{aligned} \tag{14}$$

The terms correspond to classification, mask BCE, mask Dice, box L_1, GIoU, and RFO regression, respectively. Final, auxiliary, and encoder outputs receive the corresponding set losses; the denoising branch additionally receives the RFO loss $\mathcal{L}_{\mathrm{DN}}$.

4 Experiments

4.1 Experimental Settings

Datasets. We use five named data sources. BONAI (Wang et al., 2023) supports the main comparison, IRSA (Meng et al., 2025) the ablations, and BONAI, IRSA, WHU (Ji et al., 2019), and UAV (Li et al., 2026a) the cumulative-source study; Huizhou (Li et al., 2024a) is held out from training for cross-domain evaluation.

Implementation Details. Unless stated otherwise, single-source models use an ImageNet-22K-pretrained Swin-Base backbone (Liu et al., 2021), nine decoder layers, 300 queries, and $d = 5$. We optimize with AdamW (Loshchilov & Hutter, 2019) for 35 epochs on six NVIDIA RTX 3090 GPUs. Inputs use a 768×768 canvas with scale jittering, flipping, and rotation. All ablations use the same IRSA split and training protocol.

Evaluation Metrics. Following the roof-mask/RFO protocol (Wang et al., 2023; Li et al., 2024b), bAP/sAP evaluate roof detection/instance segmentation and bAP50/sAP50 use IoU 0.5. FAP50 and best-threshold F1 evaluate offset-corrected footprints, while Boundary IoU (B-IoU) measures boundary consistency. Mean endpoint error (mEPE) is the average pixel-space RFO endpoint error; PCP@5 is the percentage below five pixels. Higher is better except for mEPE.

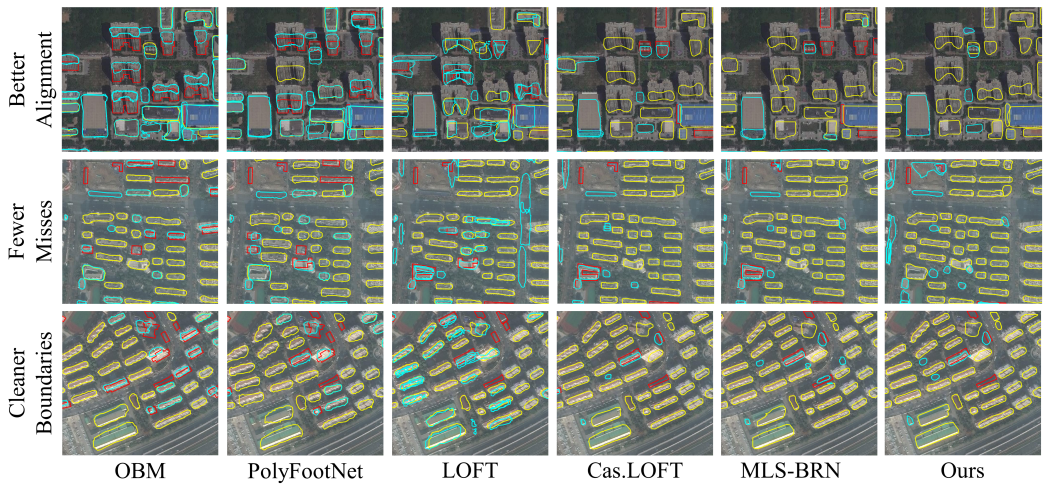


Figure 2: Qualitative comparison of OBM, PolyFootNet, LOFT, Cas.LOFT, MLS-BRN, and LoDEOT (ours) on three representative BONAI scenes. Each row shows the same scene across methods, and each column corresponds to one method. Yellow denotes true-positive matches (IoU $\geq$ 0.5), cyan unmatched predictions (false positives), and red unmatched ground-truth instances (false negatives). OBM and PolyFootNet use LoDEOT-predicted masks as guidance.

4.2 Comparison with Existing Methods

Table 1 compares LoDEOT with end-to-end LOFT (Wang et al., 2023), Cas.LOFT (Li et al., 2024a), and MLS-BRN (Li et al., 2024b) on BONAI. OBM (Li et al., 2024a) and PolyFootNet (Li et al., 2025a) use LoDEOT-predicted masks as instance guidance and serve as supplementary references. The end-to-end baselines share roof-mask and signed-RFO outputs, but architecture and training differences make this a system-level comparison rather than an isolated test of the compact offset head. Figure 2 shows all six methods on the same three scenes.

Table 1: Performance comparison of externally guided and end-to-end methods on the BONAI test set. The best results among the end-to-end methods are highlighted. All metrics are percentages except mEPE, which is measured in pixels.

Method	Roof Detection and Segmentation				Offset-Corrected Footprint				
	bAP↑	bAP50↑	sAP↑	sAP50↑	FAP50↑	F1↑	B-IoU↑	mEPE↓	PCP@5↑
Externally Guided Methods									
OBM	–	–	12.59	36.71	22.21	41.72	9.25	5.72	53.68
PolyFootNet	–	–	8.91	33.18	18.76	39.75	7.32	6.54	44.71
End-to-End Methods									
LOFT	10.50	42.90	21.00	59.50	37.73	49.64	8.78	5.65	57.74
Cas.LOFT	15.80	51.45	**30.21**	**63.25**	47.02	55.60	11.12	5.52	60.76
MLS-BRN	15.45	51.63	30.20	62.81	42.73	53.99	11.17	5.67	61.02
LoDEOT(Ours)	**31.38**	**57.20**	26.37	54.94	**54.58**	**58.87**	**11.20**	**5.23**	**64.25**

LoDEOT leads the end-to-end methods in bAP/bAP50 (31.38/57.20) and all five footprint metrics, and outperforms both externally guided references on all reported metrics. However, its sAP/sAP50 trail the roof-segmentation leader Cas.LOFT by 3.84/8.31 points, indicating stronger roof localization and RFO recovery without uniformly better roof masks (Table 1).

4.3 CUMULATIVE MULTI-SOURCE TRAINING AND CROSS-DOMAIN EVALUATION

We construct cumulative levels BONAI (L1), BONAI+IRSA (L2), BONAI+IRSA+WHU (L3), and BONAI+IRSA+WHU+UAV (L4), and evaluate each on all four sources plus unseen Huizhou. Table 2 reports sAP50/FAP50 and their unweighted five-domain averages.

Table 2: Evaluation of LoDEOT with $d = 5$ under cumulative multi-source training. The average is the unweighted mean over the five test domains.

Level	BONAI		IRSA		WHU		UAV		Huizhou		Average	
	sAP50	FAP50	sAP50	FAP50	sAP50	FAP50	sAP50	FAP50	sAP50	FAP50	sAP50	FAP50
L1	54.94	**54.58**	9.70	8.77	49.50	48.75	31.64	28.49	35.72	31.17	36.30	34.35
L2	54.68	51.58	**51.67**	**48.76**	26.65	26.60	14.03	12.98	8.82	5.56	31.17	29.09
L3	54.21	51.76	49.42	46.30	**96.08**	**96.08**	24.25	22.65	31.22	25.01	51.04	48.36
L4	**55.22**	52.29	49.27	46.34	95.55	95.54	**44.99**	**41.75**	**39.13**	**32.28**	**56.83**	**53.64**

Each added source gives its largest immediate FAP50 gain on its own domain. Performance is non-monotonic: IRSA and WHU peak at L2 and L3; BONAI peaks at L4 for sAP50 and L1 for FAP50, while UAV and unseen Huizhou peak at L4 for both. Relative to L1, L4 improves the macro-average sAP50/FAP50 by 20.53/19.29 points to 56.83/53.64 and Huizhou scores by 3.41/1.11 points. Since source composition, sample count, and training exposure vary jointly, these results describe cumulative-source effects rather than a controlled scaling law.

4.4 ABLATION STUDIES

Offset-Token Width. Table 3 evaluates 18 widths with five seeds each. $d = 7$ has the highest means, but increasing width does not monotonically improve performance. Paired two-sided t-tests comparing $d = 5$ with other widths detect no significant differences on the six metrics (raw $p \geq 0.098$; Holm correction over 102 tests). This does not establish equivalence or optimality.

The proprietary IRSA collection uses a fixed satellite-imaging modality. Under shared calibration and processing with locally restricted geometry, some physical factors can become shared or coupled; Appendix A derives these conditional reductions and their error bounds. This offers a possible explanation for similar results at $d = 2$–4, while $d = 5$ remains the default motivated by the general five-parameter organization.

Representation and Readout Cost. Relative to $d = 256$, $d = 5$ reduces offset-token coordinates by 98.05% and parameters per distinct affine readout ($2d + 2$) from 514 to 12, a 97.67% reduction (Appendix B), with mean FAP50/F1 lower by 1.06/0.79 points (Table 3).

Layer-Wise RFO Denoising. Both denoising modes improve all nine metrics over the disabled path (Table 4). Additive perturbation raises FAP50/F1 by 9.58/8.63 points and reduces mEPE from 3.29 to 2.71 pixels; relative to multiplicative perturbation, it improves FAP50/F1 by 2.58/2.30 points and reduces mEPE by 0.33 pixels. Additive leads eight metrics, with multiplicative ahead by 0.02 points in B-IoU, so additive is the default. The enabled–disabled comparison evaluates perturbation and supervision jointly, without isolating their effects.

4.5 DISCUSSION

Experimental Evidence and Operation. The comparisons evaluate complete systems and the joint denoising path; mask pooling, concentration gating, and box–mask geometry are not independently validated. Inference updates compact offset tokens from image, query, roof-mask, and box evidence; denoising groups, target indices, and lifted noisy states are training-only. Each predicted RFO applies a constant roof-mask translation without camera recovery. Whole-model cost also includes the shared encoder and decoder.

Table 3: Offset-token width ablation on the IRSA test set. Results are mean ± sample standard deviation over five random seeds.

d	bAP↑	bAP50↑	sAP↑	sAP50↑	FAP50↑	F1↑
2	29.66 ± 1.20	49.10 ± 1.56	28.66 ± 1.20	48.96 ± 1.67	45.67 ± 1.60	51.60 ± 1.20
3	29.63 ± 0.82	48.96 ± 1.42	28.73 ± 0.83	49.01 ± 1.48	45.76 ± 1.28	51.62 ± 1.34
4	29.04 ± 1.25	48.03 ± 2.04	28.16 ± 1.19	48.03 ± 2.06	45.07 ± 1.94	50.97 ± 1.55
5	29.42 ± 1.50	48.54 ± 2.35	28.41 ± 1.47	48.38 ± 2.46	45.20 ± 2.52	50.93 ± 2.16
6	30.08 ± 0.97	49.68 ± 0.74	29.10 ± 0.76	49.62 ± 0.72	46.32 ± 0.98	51.97 ± 0.69
7	$\mathbf{30.61} \pm 0.51$	$\mathbf{50.31} \pm 0.79$	$\mathbf{29.62} \pm 0.48$	$\mathbf{50.35} \pm 0.89$	$\mathbf{47.08} \pm 0.76$	$\mathbf{52.91} \pm 0.66$
8	28.74 ± 2.28	47.81 ± 3.43	27.80 ± 2.30	47.86 ± 3.28	44.46 ± 3.30	50.62 ± 2.33
9	29.60 ± 1.10	49.07 ± 2.00	28.62 ± 1.22	49.03 ± 2.13	45.87 ± 2.19	51.66 ± 1.74
10	29.66 ± 0.94	49.25 ± 1.47	28.71 ± 0.86	49.16 ± 1.43	46.02 ± 1.33	51.70 ± 1.22
11	29.58 ± 0.97	48.99 ± 1.36	28.69 ± 0.92	48.93 ± 1.33	45.86 ± 1.34	51.37 ± 1.11
12	29.91 ± 0.51	49.39 ± 0.39	29.05 ± 0.33	49.42 ± 0.31	46.33 ± 0.36	51.99 ± 0.16
13	29.89 ± 1.22	49.41 ± 1.84	28.88 ± 1.26	49.32 ± 1.96	46.20 ± 2.06	52.02 ± 1.96
14	29.56 ± 1.09	48.77 ± 1.65	28.55 ± 1.00	48.75 ± 1.65	45.58 ± 1.58	51.39 ± 1.41
15	29.61 ± 1.37	48.92 ± 2.31	28.64 ± 1.47	48.88 ± 2.29	45.71 ± 2.04	51.25 ± 1.80
16	29.79 ± 0.45	49.19 ± 0.79	28.94 ± 0.32	49.29 ± 0.80	46.02 ± 0.67	51.74 ± 0.67
32	29.68 ± 0.86	49.22 ± 1.28	28.87 ± 0.79	49.32 ± 1.27	46.07 ± 1.28	51.56 ± 1.21
64	30.12 ± 0.85	49.54 ± 1.63	29.15 ± 0.96	49.55 ± 1.61	46.66 ± 1.39	52.23 ± 1.32
256	29.92 ± 1.05	49.47 ± 1.64	28.95 ± 1.07	49.47 ± 1.64	46.26 ± 1.63	51.72 ± 1.24

Table 4: Ablation of the target-indexed RFO denoising path on the IRSA test set with $d = 5$. Multiplicative and additive denote perturbations applied to the initial offset token.

Configuration	bAP↑	bAP50↑	sAP↑	sAP50↑	FAP50↑	F1↑	B-IoU↑	mEPE↓	PCP@5↑
Disabled	24.02	40.59	23.06	40.57	37.21	43.91	21.13	3.29	84.58
Multiplicative	28.33	47.32	27.49	47.66	44.21	50.24	**22.29**	3.04	85.52
Additive	**30.27**	**49.57**	**29.17**	**49.65**	**46.79**	**52.54**	22.27	**2.71**	**88.09**

Dimensionality and Scope. The five-factor organization motivates $d = 5$ without fixed physical meanings for learned channels; restricted acquisition conditions can explain smaller widths' effectiveness. The factorization assumes local pinhole projection and vertical extrusion. Within-tile displacement directions can vary in UAV imagery (Li et al., 2026a), limiting a shared coefficient and motivating broader width comparisons. Uncertain roof masks can weaken pooled appearance and centroid cues, so better footprints need not imply better roof segmentation.

Broader Implications. LoDEOT addresses geometric alignment in building instance segmentation, separating appearance and displacement representations to connect roof delineation with footprint localization through a compact design.

5 Conclusion

We derive an exact five-parameter RFO factorization under local pinhole projection and vertical extrusion, providing a photogrammetric basis for compact offset-token design. Guided by this analysis, LoDEOT combines a mask-guided five-dimensional offset token with target-indexed layer-wise denoising while preserving high-dimensional multitask instance tokens. LoDEOT leads the compared end-to-end methods on all five BONAI footprint metrics; denoising improves all ablation metrics, and the largest cumulative source mixture achieves the best five-domain macro-average. Width comparisons detect no significant differences from $d = 5$; conditional factor sharing and coupling may explain similar performance at smaller widths. Together, these findings support separating appearance modeling and geometric displacement prediction through compact offset representations and readouts for building instance segmentation.

AI USE STATEMENT

In preparing this manuscript, we used generative AI tools to assist with English writing and translation, check technical terminology, and improve clarity and readability. The authors reviewed and revised the AI-assisted text and take responsibility for the final manuscript.

REFERENCES

David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient interactive annotation of segmentation datasets with Polygon-RNN++. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 859–868, 2018.

Yeshwanth Kumar Adimoolam, Charalambos Poullis, and Melinos Averkiou. Pix2Poly: A sequence prediction method for end-to-end polygonal building footprint extraction from remote sensing imagery. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 8473–8482, 2025.

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. AC3D: Analyzing and improving 3D camera control in video diffusion transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22875–22889. IEEE, 2025.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pp. 213–229, 2020. doi: 10.1007/978-3-030-58452-8_13.

Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17864–17875, 2021.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, 2022.

Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. DeepGlobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 172–181, 2018.

Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6910–6919, 2021.

Nicolas Girard, Dmitriy Smirnov, Justin Solomon, and Yuliya Tarabalka. Polygonal building extraction by frame field learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5891–5900, 2021.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.

Yuan Hu, Zhibin Wang, Zhou Huang, and Yu Liu. PolyBuilding: Polygon transformer for building extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 199:15–27, 2023. doi: 10.1016/j.isprsjprs.2023.03.021.

Zhenghao Hu, Weijia Li, Jinhua Yu, Minfa Liu, Junyan Ye, Peimin Chen, and Huabing Huang. 3D building reconstruction from monocular remote sensing imagery via diffusion models and geometric priors. *ISPRS Journal of Photogrammetry and Remote Sensing*, 232:124–137, 2026. doi: 10.1016/j.isprsjprs.2025.11.018.

Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2019. doi: 10.1109/TGRS.2018.2858817.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. DN-DETR: Accelerate DETR training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13619–13627, 2022.

Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask DINO: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3041–3050, 2023.

Kai Li, Yupeng Deng, Yunlong Kong, Diyou Liu, Jingbo Chen, Yu Meng, Junxian Ma, and Chenhao Wang. Prompt-driven building footprint extraction in aerial images with offset-building model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024a. doi: 10.1109/TGRS.2024.3487652.

Kai Li, Yupeng Deng, Jingbo Chen, Yu Meng, Zhihao Xi, Junxian Ma, Chenhao Wang, Maolin Wang, and Xiangyu Zhao. PolyFootNet: Extracting polygonal building footprints in off-nadir remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–16, 2025a. doi: 10.1109/TGRS.2025.3590054. Art. no. 5635016.

Kai Li, Xingxing Weng, Yupeng Deng, Yu Meng, Chao Pang, Gui-Song Xia, and Xiangyu Zhao. DragOSM: Extract building roofs and footprints from aerial images by aligning historical labels, 2025b. URL https://arxiv.org/abs/2509.17951.

Kai Li, Yupeng Deng, Ligao Deng, Zhihao Xi, Chenhao Wang, Jierui Zhang, Yingrui Ji, Yu Meng, and Xiangyu Zhao. ObliCity: A benchmark and baseline for roof-to-ground projection displacement correction. *arXiv preprint arXiv:2607.25210*, 2026a.

Weijia Li, Wenqian Zhao, Huaping Zhong, Conghui He, and Dahua Lin. Joint semantic-geometric learning for polygonal building segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):1958–1965, 2021. doi: 10.1609/aaai.v35i3.16291.

Weijia Li, Haote Yang, Zhenghao Hu, Juepeng Zheng, Gui-Song Xia, and Conghui He. 3D building reconstruction from monocular remote sensing images with multi-level supervisions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27728–27737, 2024b.

Yuxuan Li, Keming Chen, Zicheng Lei, Yating Yang, Jinsong Lv, and Zhi Zhou. RBF-SAM: Robust building footprint extraction from off-nadir remote sensing images via segment anything model. *IEEE Transactions on Geoscience and Remote Sensing*, 64, 2026b. doi: 10.1109/TGRS.2026.3709903. Art. no. 5631416.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the Inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3226–3229, 2017. doi: 10.1109/IGARSS.2017.8127684.

Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional DETR for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3651–3660, 2021.

Yu Meng, Ligao Deng, Zhihao Xi, Jiansheng Chen, Jingbo Chen, Anzhi Yue, Diyou Liu, Kai Li, Chenhao Wang, Kaiyu Li, Yupeng Deng, and Xian Sun. IRSAMap: Toward large-scale, high-resolution land cover map vectorization. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–19, 2025. doi: 10.1109/TGRS.2025.3600249.

Bipul Neupane, Jagannath Aryal, Abbas Rajabifard, Patrick Aravena Pelizari, and Christian Geiß. Multilabel learning with ViT for building footprint extraction from off-nadir aerial images. *IEEE Geoscience and Remote Sensing Letters*, 22:1–5, 2025. doi: 10.1109/LGRS.2025.3532589. Art. no. 8000605.

Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14356–14366. IEEE, 2021.

Jinwang Wang, Lingxuan Meng, Weijia Li, Wen Yang, Lei Yu, and Gui-Song Xia. Learning to extract building footprints from off-nadir aerial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1294–1301, 2023. doi: 10.1109/TPAMI.2022.3162583.

Shiqing Wei, Tao Zhang, Shunping Ji, Muying Luo, and Jianya Gong. BuildMapper: A fully learnable framework for vectorized building contour extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:87–104, 2023. doi: 10.1016/j.isprsjprs.2023.01.015.

Bowen Xu, Jiakun Xu, Nan Xue, and Gui-Song Xia. HiSup: Accurate polygonal mapping of buildings in satellite imagery with hierarchical supervision. *ISPRS Journal of Photogrammetry and Remote Sensing*, 198:284–296, 2023. doi: 10.1016/j.isprsjprs.2023.03.006.

Ziqiang Xu, Chunyan Xu, Zhen Cui, Xiangwei Zheng, and Jian Yang. CVNet: Contour vibration network for building extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1383–1391, 2022.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2023.

Weichen Zhou, Yawen Zou, Chunzhi Gu, Ran Dong, Haoran Xie, and Chao Zhang. Understanding geometric representations in self-supervised vision transformers via subspace intervention. In *European Conference on Computer Vision*, pp. 357–374. Springer, 2026.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

Stefano Zorzi, Shabab Bazrafkan, Stefan Habenschuss, and Friedrich Fraundorfer. PolyWorld: Polygonal building extraction with graph neural networks in satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1848–1857, 2022.

A Conditional Reduction of the RFO Map under Restricted Imaging Geometry

This appendix derives the local approximation $\boldsymbol{\delta} \simeq H\mathbf{k}_0$ to help interpret the width study in Section 4.4 and bounds its approximation error. The analysis uses the local pinhole and vertical-extrusion assumptions of Section 3.2. It distinguishes shared camera parameters, weak height-dependent perspective effects, and locally restricted geometry; these conditions need not hold uniformly across a dataset. Vector norms are Euclidean and matrix norms are the induced spectral norms.

A.1 Exact Offset and Shared Calibration

Let $(X, Y, Z)^\top$ be the camera-frame footprint point and $\boldsymbol{\nu} = (\nu_x, \nu_y, \nu_z)^\top \in \mathbb{S}^2$ the camera-frame vertical direction. A roof reference point has camera coordinates $(X, Y, Z)^\top + H\boldsymbol{\nu}$, where $H > 0$. Require $Z > 0$ and $Z + H\nu_z > 0$. Write $\mathbf{x} = (X/Z, Y/Z)^\top$, $\boldsymbol{\nu}_{xy} = (\nu_x, \nu_y)^\top$, and

$$\mathbf{A} = \begin{bmatrix} f_x & \kappa \\ 0 & f_y \end{bmatrix} = f_y \begin{bmatrix} a & \widetilde{\kappa} \\ 0 & 1 \end{bmatrix}, \qquad \boldsymbol{\beta} = \nu_z \mathbf{x} - \boldsymbol{\nu}_{xy}. \tag{A.1}$$

Here $f_x, f_y > 0$, $a = f_x/f_y$, and $\widetilde{\kappa} = \kappa/f_y$, consistently with Section 3.2. The principal point cancels when subtracting the two projections, giving

$$\begin{aligned} \boldsymbol{\delta} &= \mathbf{A}\left(\mathbf{x} - \frac{Z\mathbf{x} + H\boldsymbol{\nu}_{xy}}{Z + H\nu_z}\right) \\ &= \frac{H}{Z + H\nu_z}\mathbf{A}\boldsymbol{\beta} = \frac{H/Z}{1 + (H/Z)\nu_z}\mathbf{A}\boldsymbol{\beta}. \end{aligned} \tag{A.2}$$

This is the signed roof-to-footprint displacement in Eq. (5). Under a shared effective calibration and processing regime, a and $\widetilde{\kappa}$ are shared across instances. For the local bound below, we hold the full matrix $\mathbf{A}$ fixed within a neighborhood; different neighborhoods or acquisitions may have different matrices. A fixed imaging modality or a standardized ground sampling distance alone does not establish this calibration condition.

A.2 Height-Dependent Perspective Remainder

Define $\rho = H/Z$, $\tau = \rho\nu_z$, and the displacement obtained by omitting the height-dependent denominator, $\boldsymbol{\delta}_{\mathrm{lin}} = \rho\mathbf{A}\boldsymbol{\beta}$. Equation (A.2) gives the exact identity

$$\begin{aligned} \boldsymbol{\delta} - \boldsymbol{\delta}_{\mathrm{lin}} &= -\tau\boldsymbol{\delta}, \\ \frac{\|\boldsymbol{\delta} - \boldsymbol{\delta}_{\mathrm{lin}}\|}{\|\boldsymbol{\delta}\|} &= |\tau| \le \rho \qquad (\boldsymbol{\delta} \neq \mathbf{0}). \end{aligned} \tag{A.3}$$

Thus, small H/Z controls this perspective correction, independently of the focal scale. For a uniform bound, suppose $H \le H_{\max}$ and $Z \ge Z_{\min} > 0$, and set $\eta = H_{\max}/Z_{\min} < 1$. Then $|\tau| \le \eta$ and $1 + \tau \ge 1 - \eta$, so

$$\|\boldsymbol{\delta} - \boldsymbol{\delta}_{\mathrm{lin}}\| \le \frac{\eta}{1 - \eta}\|\boldsymbol{\delta}_{\mathrm{lin}}\|. \tag{A.4}$$

This absolute bound also covers zero displacement. It controls the approximation within the assumed local camera model, not errors from replacing a physical imaging system by that model.

A.3 Local Variation of the Geometric Factor

Choose a reference state $(Z_0, \mathbf{x}_0, \boldsymbol{\nu}_0)$ with $Z_0 > 0$, and define $\boldsymbol{\beta}_0 = \nu_{0z}\mathbf{x}_0 - \boldsymbol{\nu}_{0,xy}$. The reference subscripts denote components of $\boldsymbol{\nu}_0$. Direct subtraction gives

$$\begin{aligned} \boldsymbol{\beta} - \boldsymbol{\beta}_0 &= \nu_z(\mathbf{x} - \mathbf{x}_0) + (\nu_z - \nu_{0z})\mathbf{x}_0 - (\boldsymbol{\nu}_{xy} - \boldsymbol{\nu}_{0,xy}), \\ \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| &\le \|\mathbf{x} - \mathbf{x}_0\| + (\|\mathbf{x}_0\| + 1)\|\boldsymbol{\nu} - \boldsymbol{\nu}_0\|. \end{aligned} \tag{A.5}$$

The inequality follows from $|\nu_z| \le 1$ and the triangle inequality. If $\|\mathbf{x} - \mathbf{x}_0\| \le \varepsilon_x$ and $\|\boldsymbol{\nu} - \boldsymbol{\nu}_0\| \le \varepsilon_\nu$, then $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \varepsilon_\beta$, where $\varepsilon_\beta = \varepsilon_x + (\|\mathbf{x}_0\| + 1)\varepsilon_\nu$. For exactly shared viewing orientation, the difference reduces to $\nu_z(\mathbf{x} - \mathbf{x}_0)$; small normalized spatial extent therefore limits geometric variation even when building heights differ.

A.4 A Shared Local Coefficient and Its Error Bound

Assume in addition that $|Z - Z_0| \le \varepsilon_Z Z_0$ with $0 \le \varepsilon_Z < 1$. Define the locally shared coefficient vector $\mathbf{k}_0 = \mathbf{A}\boldsymbol{\beta}_0/Z_0$. It has units of image displacement per unit building height. Adding and subtracting $\mathbf{A}\boldsymbol{\beta}_0/Z$ gives

$$\begin{aligned} \frac{\mathbf{A}\boldsymbol{\beta}}{Z} - \mathbf{k}_0 &= \frac{\mathbf{A}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)}{Z} + \mathbf{A}\boldsymbol{\beta}_0 \left(\frac{1}{Z} - \frac{1}{Z_0}\right), \\ \left\| \frac{\mathbf{A}\boldsymbol{\beta}}{Z} - \mathbf{k}_0 \right\| &\le \frac{\|\mathbf{A}\|}{Z_0(1-\varepsilon_Z)} \left(\varepsilon_\beta + \varepsilon_Z \|\boldsymbol{\beta}_0\|\right). \end{aligned} \tag{A.6}$$

Indeed, $Z \ge Z_0(1-\varepsilon_Z)$ and $|1/Z - 1/Z_0| \le \varepsilon_Z/[Z_0(1-\varepsilon_Z)]$. Since $\boldsymbol{\delta}_{\text{lin}} = H\mathbf{A}\boldsymbol{\beta}/Z$, the residual $\mathbf{e} = \boldsymbol{\delta} - H\mathbf{k}_0$ is the sum of $\boldsymbol{\delta} - \boldsymbol{\delta}_{\text{lin}}$ and $H(\mathbf{A}\boldsymbol{\beta}/Z - \mathbf{k}_0)$. Combining Eqs. (A.4) and (A.6), with $\|\boldsymbol{\beta}\| \le \|\boldsymbol{\beta}_0\| + \varepsilon_\beta$, yields

$$\begin{aligned} \boldsymbol{\delta} &= H\mathbf{k}_0 + \mathbf{e}, \\ \|\mathbf{e}\| &\le \frac{H\|\mathbf{A}\|}{Z_0(1-\varepsilon_Z)} \left[\frac{\eta}{1-\eta} \left(\|\boldsymbol{\beta}_0\| + \varepsilon_\beta\right) + \varepsilon_\beta + \varepsilon_Z \|\boldsymbol{\beta}_0\| \right]. \end{aligned} \tag{A.7}$$

The three terms account for height-dependent perspective, local geometric variation, and local depth variation. Their smallness gives a controlled absolute approximation; when the leading displacement is nearly zero, it need not give a small relative error. For $\mathbf{k}_0 \neq \mathbf{0}$, the leading displacements lie in the one-dimensional subspace spanned by $\mathbf{k}_0$, and $|k_{0y}\delta_x - k_{0x}\delta_y| \le \|\mathbf{k}_0\|\|\mathbf{e}\|$. This relation avoids dividing by a possibly zero displacement component. If $\mathbf{k}_0 = \mathbf{0}$, Eq. (A.7) still bounds the displacement, but no common direction is defined.

Directional coupling and linear dependence on height are distinct. If $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ exactly and calibration is shared, Eq. (A.2) already gives $\boldsymbol{\delta} = [HZ_0/(Z + H\nu_z)]\mathbf{k}_0$: both components share one scalar amplitude even without a weak-perspective approximation. Small H/Z and small relative depth variation further make this amplitude approximately H. The coefficient is local; the argument does not require one coefficient vector for an entire image or dataset.

The bound also admits approximate calibration sharing. If $\|\mathbf{A} - \mathbf{A}_0\| \le \varepsilon_A$, define $\mathbf{k}_0 = \mathbf{A}_0\boldsymbol{\beta}_0/Z_0$ and use $\mathbf{A}_0$ in Eq. (A.7). The additional residual equals $H(\mathbf{A} - \mathbf{A}_0)\boldsymbol{\beta}/[Z(1+\tau)]$ and is bounded by $H\varepsilon_A(\|\boldsymbol{\beta}_0\| + \varepsilon_\beta)/[Z_{\min}(1-\eta)]$. Thus small effective-calibration variation contributes an explicit additional error term rather than requiring exact equality.

A.5 Satellite Evidence and Interpretation of the Width Study

The proprietary IRSA collection used here has a fixed satellite-imaging modality. The related IRSAMap documentation reports that its Chinese imagery was selected from 13 Jilin-1 scenes and standardized to a ground sampling distance of $0.5\,\text{m}$ (Meng et al., 2025). This documents a setting in which many image crops originate from a finite set of acquisitions: additional crops need not supply independently varying camera conditions. These public provenance details are contextual evidence and do not specify the scene inventory or calibration of the proprietary collection.

The "Jilin-1 GF03D" technical specifications on the official website of Chang Guang Satellite Technology Co., Ltd. (accessed September 26, 2026) list an orbital altitude of $535\,\text{km}$, pushbroom imaging, and a side-sway range of $\pm 45^\circ$. The orbital scale motivates examining $H/Z \ll 1$. For illustration, taking an optical depth $Z = 535\,\text{km}$ and a building height $H = 300\,\text{m}$ gives $H/Z \approx 5.61 \times 10^{-4}$, so Eq. (A.3) bounds the relative height-perspective correction by approximately 0.056%. This is an illustrative calculation, not a measured IRSA bound: orbital altitude is not generally the optical-axis depth Z, and the cited specifications do not identify the satellite subtype used for each proprietary image. The pointing range also precludes inferring a common viewing direction across acquisitions from the satellite name alone.

All derivations are conditional on the local pinhole approximation. Spatially varying pushbroom geometry, orthorectification, or mosaicking may change the effective local camera map, and their modeling errors are not included in Eq. (A.7). A common affine image transformation preserves displacement collinearity and maps the residual to the same linear transformation of $\mathbf{e}$; spatially

varying transformations require separate local analysis. In particular, ObliCity reports varying displacement directions within UAV tiles due to low-altitude acquisition and mosaicking (Li et al., 2026a), illustrating why one shared coefficient may fail over a larger image region.

The general descriptor retains intrinsic shape, displacement amplitude, and relative geometry as separate factors. Shared calibration and restricted local geometry can reduce their independent variation in a particular acquisition setting, offering a physical interpretation of similar performance at smaller widths. We therefore retain $d = 5$ as the default motivated by the general factor organization, without assigning physical identities to its learned channels. These conditional results neither establish a minimum neural width nor identify the cause of non-significant width differences; broader acquisition diversity, rather than dataset openness itself, determines where the local simplifications should be reassessed.

B Offset Representation and Readout Parameters

For K instances and offset-token width d, the offset representation contains Kd scalars. The affine readout $\hat{\mathbf{z}} = \mathbf{W}\mathbf{o} + \mathbf{b}_{\text{out}}$, where $\mathbf{W} \in \mathbb{R}^{2\times d}$ and $\mathbf{b}_{\text{out}} \in \mathbb{R}^2$, contains $2d + 2$ trainable parameters, including the bias. Table B.1 reports these quantities for selected widths. Counts are per token collection and per distinct readout module; shared readout parameters are counted once. These quantities cover the offset representation and its readout, while the update modules and shared backbone and decoder contribute additional parameters.

Table B.1: Offset-representation size and affine-readout parameters, including bias, for $K = 300$.

d	Token scalars Kd	Readout parameters $2d + 2$
5	1,500	12
7	2,100	16
256	76,800	514

Relative to $d = 256$, $d = 5$ reduces token coordinates by $1 - 5/256 \approx 98.05\%$ and readout parameters by $1 - 12/514 \approx 97.67\%$. The reported parameter reduction applies to each distinct affine readout, including its bias, and does not describe a full-model parameter reduction.