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

Anonymous submission

Abstract

Instance-level roof-to-footprint offset (RFO) prediction is central to extracting building footprints from off-nadir imagery. Query-based pipelines usually carry a generic high-dimensional instance embedding to an offset terminal whose output is only a signed two-dimensional displacement. We study whether this task-specific terminal can instead be compact. Under local pinhole projection and vertical-extrusion assumptions, the idealized RFO map admits a seven-variable sufficient descriptor. We use this factorization only to motivate a candidate neural width: learned token coordinates remain latent and are not assigned physical meanings or asserted to be equivalent to the physical variables. Based on this distinction, we propose LoDEOT, which retains high-dimensional queries for detection and segmentation but maps decoder-query, quality-gated roof, and box–mask evidence to a low-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. On BONAI, LoDEOT achieves the best roof-detection bAP and bAP50 among the end-to-end methods and leads all five offset-corrected footprint metrics, with FAP50 reaching 49.91 and mEPE reduced to 5.24 pixels. A 17-setting width sweep supports seven as a favorable performance–terminal-complexity choice under the evaluated single-seed protocol, without claiming uniqueness or universal optimality.

1 Introduction

Accurate building footprints support urban mapping, population analysis, disaster response, and three-dimensional city modeling. Raster segmentation and polygonal extraction have substantially improved building delineation (Maggiori et al. 2017; Demir et al. 2018; Ji, Wei, and Lu 2019; Li et al. 2021; Girard et al. 2021), but off-nadir imagery introduces a distinct geometric ambiguity: viewing direction and building height displace a visible roof from its ground outline in the image plane. A roof mask alone is therefore not an accurate footprint. The signed roof-to-footprint offset (RFO) explicitly links the projected roof to the footprint, directly affects footprint localization and boundary quality, and also supports height estimation, LoD2-level modeling, and map correction (Wang et al. 2022; Li et al. 2024b).

RFO prediction has progressed from convolutional roof–offset branches (Wang et al. 2022) to prompt-driven and query-based instance models (Li et al. 2024a,b, 2025). These methods improve observations, supervision, or output structure, yet the terminal representation used to predict the offset itself remains largely unexamined, *e.g.,* commonly used 256, 512, 1024 dimensions for token design. A generic instance query must stay expressive for classification, localization, and segmentation; forcing the task-specific offset route to inherit the same width adds width-dependent terminal parameters and leaves roof appearance, mask reliability, and box–mask asymmetry only implicitly available. The issue is not whether to compress the shared query, but whether the RFO route can be decoupled into a compact terminal without reducing the capacity of the other heads.

Two observations guide our design. First, under local pinhole-camera and vertical-extrusion assumptions, principal-point cancellation, coordinate gauges, and scale coupling factor the idealized RFO map through seven variables. This result is a width-selection prior, not an identification theorem for a neural representation: the learned channels need not correspond to physical variables, and seven is not asserted to be unique, minimal, or universally optimal. Second, a Transformer decoder produces successive instance-conditioned estimates. Denoising queries retain target indices under label/box perturbations (Li et al. 2022; Zhang et al. 2023; Li et al. 2023); those indices can align every supervised layer with the same clean RFO without another Hungarian assignment.

We consequently propose LoDEOT. High-dimensional decoder queries continue to serve the class, roof-mask, and roof-box heads. At each layer, a mask-guided OffsetHead combines the current query with quality-gated position-aware roof evidence and box–mask spatial cues, maps this evidence to a compact token, and uses a separate linear readout for the signed two-dimensional RFO. The predicted vector translates the roof mask into the footprint. During training, target-indexed denoising supervision binds successive estimates to the same clean instance offset, giving the layer-wise outputs an error-recovery interpretation without recurrently propagating a single offset token.

Experiments examine the design from four complementary views. On BONAI (Wang et al. 2022), LoDEOT achieves the best roof-detection bAP and bAP50 scores among the end-to-end methods and leads all five offset-corrected footprint metrics, reaching 49.91 FAP50 and a 5.24-pixel mEPE. Across 17 tested widths, $d = 7$ ties for the best Footprint AP50, gives the best F1, and uses 46.7% of the width-dependent terminal parameters of $d = 15$. Target-indexed denoising improves every reported ablation metric, while the largest cumulative source set gives the best five-domain macro-average. These results support a favorable candidate under the reported protocols rather than a universal optimum or causal scaling law.

The contributions of this work are threefold:

1. We analyze the idealized RFO map and obtain an exact seven-variable sufficient descriptor under stated imaging assumptions, using it only to propose a geometry-inspired candidate width for the neural offset terminal.
2. We introduce LoDEOT, preserving high-dimensional multitask queries while decoupling RFO prediction into a compact, mask-guided terminal; the width sweep quantifies its performance–terminal-complexity trade-off.
3. We formulate target-indexed layer-wise RFO denoising, using known query–target identities to supervise successive recovery of the same clean instance offset without additional matching.

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, Wei, and Lu 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, Poullis, and Averkiou 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. 2022); 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. 2025). 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 of the offset terminal 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. 2020; Liu et al. 2022). MaskFormer uses mask classification, and Mask2Former restricts attention to predicted mask regions (Cheng, Schwing, and Kirillov 2021; Cheng et al. 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 regenerates a compact offset token from current evidence at each layer and uses predetermined denoising identities to supervise the same clean RFO consistently.

3 Method

LoDEOT (Low-Dimensional and Efficient Offset Tokens) jointly predicts roof instances and their signed roof-to-footprint offsets (RFOs). Section 3.1 defines the task and architecture, Section 3.2 develops the physical factorization used only to propose a candidate token width, Section 3.3 constructs the mask-guided compact terminal, and Section 3.4 explains the denoising motivation, target-indexed layer-wise RFO supervision and the learning objective.

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:

$$\begin{aligned} \widehat{\mathcal{Y}} &= \{(\hat{c}_q, \hat{\mathbf{b}}_q, \hat{\mathbf{m}}_q^r, \hat{\boldsymbol{\delta}}_q)\}_{q=1}^{K}, \\ \boldsymbol{\delta}_n &= \mathbf{u}_b(n) - \mathbf{u}_r(n). \end{aligned} \tag{1}$$

Here $\hat{c}_q$, $\hat{\mathbf{b}}_q$, and $\hat{\mathbf{m}}_q^r$ are the class, roof box, and roof mask, while $\hat{\boldsymbol{\delta}}_q \in \mathbb{R}^2$ points from the projected roof reference point $\mathbf{u}_r$ to the footprint reference point $\mathbf{u}_b$. Thus, translating the predicted roof contour by the positive RFO yields the footprint. The shared queries remain high dimensional for the multitask heads; only OffsetHead maps current instance evidence through a configurable low-dimensional terminal.

Figure 1 separates these two representation roles. The image encoder and L-layer decoder retain the capacity needed to distinguish instances and support class, box, and mask

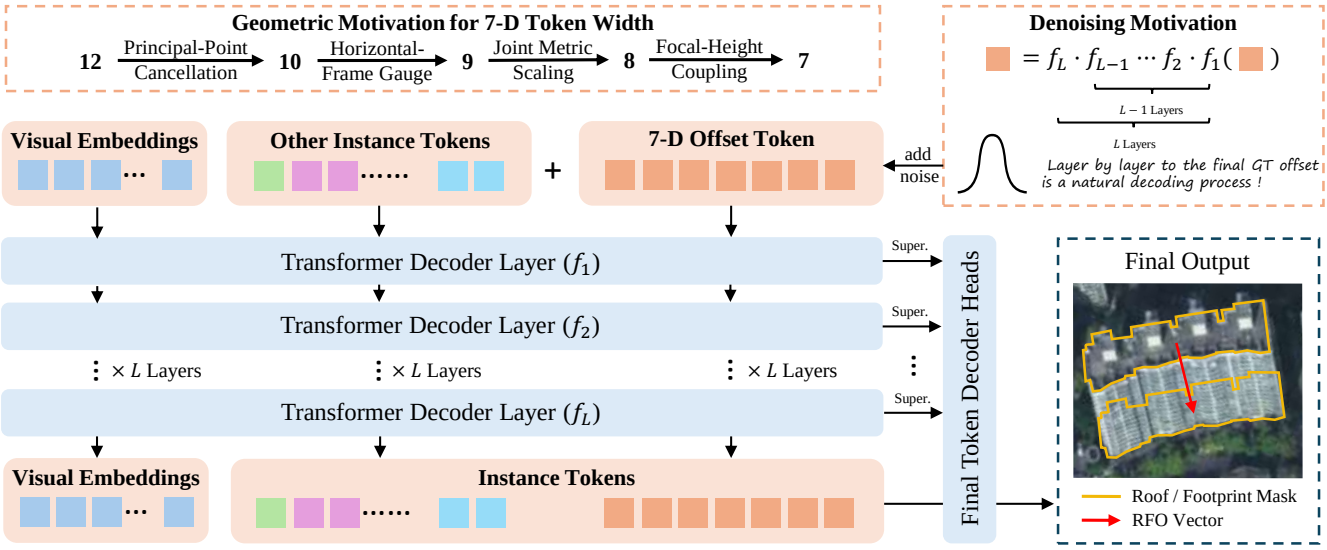

Figure 1: LoDEOT architecture. High-dimensional decoder queries feed the classification, roof-mask, and roof-box heads. At each layer, OffsetHead fuses the decoder-query feature, a quality-gated roof feature, and a box–mask spatial feature, maps them to a compact token, and reads out the signed 2-D RFO used to shift the roof mask. The upper-left path motivates seven as a candidate token width; this geometric motivation provides a width-selection prior rather than a one-to-one correspondence between physical variables and learned channels. During training, perturbed denoising queries retain their instance indices, allowing per-layer RFO estimates to share the same clean RFO target.

prediction. OffsetHead is attached to each decoder layer and compresses only the evidence used by the final two-coordinate RFO readout. Consequently, the width analysis below concerns this task-specific terminal rather than the dimensionality of the shared decoder or of the entire detector.

3.2 Geometry-Inspired Candidate Width

Off-nadir projection displaces the visible roof from the building footprint. The displacement depends jointly on camera intrinsics, relative camera–building pose, and building height, but not every coordinate of that physical description remains observable in the image-plane difference. We factor the forward map to identify a compact physically motivated candidate width; the learned terminal is then evaluated independently by the width sweep in Section 4.4.

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 world-vertical direction, and $H > 0$ the building height. Vertical extrusion gives the roof point $\mathbf{B} + H\mathbf{g}$, while the relative camera–building vector $\mathbf{r} = \mathbf{B} - \mathbf{C}$ removes the arbitrary world origin. Let $\mathbf{R} \in \mathrm{SO}(3)$ map world directions to the camera frame. A general pinhole intrinsic matrix $\mathbf{K}$ contains focal scales f_x, f_y, 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

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

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

Seven-variable factorization of the physical map. Define camera-frame vectors $\boldsymbol{\xi} = \mathbf{R}\mathbf{r} = (X, Y, Z)^\top$ and $\boldsymbol{\nu} = \mathbf{R}\mathbf{g} = (\nu_x, \nu_y, \nu_z)^\top \in \mathbb{S}^2$, normalized coordinates $x = X/Z$, $y = Y/Z$, and relative height $\rho = H/Z$. For visible roof and footprint points, $Z > 0$ and $Z + H\nu_z > 0$. Subtracting their projections yields

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

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

The principal point cancels from the difference, reducing twelve coordinates to ten. The remaining pose enters only through $\mathbf{R}\mathbf{r}$ and $\mathbf{R}\mathbf{g}$: for any rotation $\mathbf{Q}$ that preserves $\mathbf{g}$, $(\mathbf{R}, \mathbf{r}) \mapsto (\mathbf{R}\mathbf{Q}^\top, \mathbf{Q}\mathbf{r})$ leaves both vectors unchanged, so a one-dimensional horizontal-frame gauge reduces the count to nine. Joint scaling $(\mathbf{r}, H) \mapsto (\lambda\mathbf{r}, \lambda H)$ preserves (x, y, ρ) and removes absolute metric scale, leaving eight.

The last reduction couples common focal scale and relative height. With focal anisotropy $a = f_x/f_y$, normalized skew $\widetilde{\kappa} = \kappa/f_y$, and composite magnitude $\gamma = f_y\rho/(1 + \rho\nu_z)$,

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

Equivalently, for any $\mu > 0$ satisfying $\mu + (\mu - 1)\rho\nu_z > 0$, setting $\mathbf{A}' = \mu\mathbf{A}$ for $\mathbf{A} = \left[\begin{smallmatrix} f_x & \kappa \\ 0 & f_y \end{smallmatrix}\right]$ and $\rho' = \rho/[\mu + (\mu - 1)\rho\nu_z]$ preserves the offset because

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

Thus the common focal scale and ρ contribute only through γ. If (θ, φ) is a nonsingular local coordinate chart for $\boldsymbol{\nu} \in \mathbb{S}^2$, the exact dimensional reduction and a sufficient seven-variable descriptor on the visibility domain are

$$\begin{aligned} 12 &\xrightarrow{-2 \text{ principal point}} 10 \xrightarrow{-1 \text{ horizontal gauge}} 9, \\ 9 &\xrightarrow{-1 \text{ metric scale}} 8 \xrightarrow{-1 \text{ focal–height}} 7, \\ \boldsymbol{\eta}_7 &= (a, \widetilde{\kappa}, x, y, \gamma, \theta, \varphi), \qquad \boldsymbol{\delta} = \Phi_7(\boldsymbol{\eta}_7). \end{aligned} \tag{7}$$

In particular, $\boldsymbol{\eta}_7 \in \mathcal{D}_7 \subset \mathbb{R}^7$ is sufficient for the signed physical RFO under the stated local pinhole-camera, vertical-extrusion, and visibility assumptions. Sufficiency here characterizes the physical forward map: it does not make the descriptor identifiable from one image, nor does it establish that a neural predictor requires exactly seven channels.

Relation between physical and neural routes. The physical route $\boldsymbol{\eta}_7 \mapsto \boldsymbol{\delta}$ is an explicit image-formation factorization. The neural route maps image/query evidence to a learned token $\mathbf{o}_q^{(\ell)}$ and to $\hat{\boldsymbol{\delta}}_q^{(\ell)}$. They share only the signed 2-D output. We therefore use seven as a geometry-inspired candidate width and do not assign learned coordinates physical meanings or claim equivalence, identifiability, uniqueness, or minimality.

3.3 Mask-Guided Compact Offset Terminal

Position-aware roof evidence. The mask feature map $\mathbf{F}_{ij} \in \mathbb{R}^{d_m}$ is bilinearly resized to at most $S \times S$ and augmented with sinusoidal position encoding. For query q, detached mask logits are passed through a sigmoid and normalized, after which the attention pools roof-region features:

$$\begin{aligned} \widetilde{\mathbf{F}}_{ij} &= \mathbf{F}_{ij} + \mathrm{PE}_{\sin}(i, j), \\ A_{qij} &= \frac{\sigma(\mathrm{sg}(\hat{m}_{qij}))}{\sum_{i',j'} \sigma(\mathrm{sg}(\hat{m}_{qi'j'})) + \epsilon}, \\ \mathbf{p}_q &= \sum_{ij} A_{qij} \widetilde{\mathbf{F}}_{ij}, \qquad \mathbf{f}_q = f_{\mathrm{feat}}(\mathbf{p}_q). \end{aligned} \tag{8}$$

The stop-gradient operator prevents the offset loss from changing mask logits while leaving the feature branch trainable. This separation avoids using RFO regression as an indirect mask-logit objective, while still allowing the pooled feature representation to adapt. During training, a randomly sampled query subset receives zero-mean Gaussian perturbations on the unnormalized mask weights, followed by clipping to $[0, 1]$, to simulate incomplete roofs.

Mask quality and box–mask geometry. For actual pooling size $H_s \times W_s$, attention concentration defines a scalar reliability score, and the same attention distribution gives the normalized roof centroid:

$$\begin{aligned} s_q &= \max_{ij} A_{qij} \exp\left[-\frac{\mathcal{H}(\mathbf{A}_q)}{H_s W_s}\right], \\ \mathcal{H}(\mathbf{A}_q) &= -\sum_{ij} A_{qij} \log(A_{qij} + \epsilon), \\ (\bar{x}_q, \bar{y}_q) &= \sum_{ij} A_{qij}(x_{ij}, y_{ij}). \end{aligned} \tag{9}$$

The maximum attention weight favors a concentrated response, whereas the spatially scaled entropy term penalizes diffuse support. Multiplying them makes the roof feature and the centroid asymmetry less influential when the inferred roof region is spatially uncertain; it does not introduce a separate quality-prediction target. For detached reference box $\hat{\mathbf{b}}_q = (b_x^q, b_y^q, b_w^q, b_h^q)$, the quality-weighted centroid asymmetry and spatial feature are

$$\begin{aligned} \mathbf{d}_q^{\mathrm{bm}} &= s_q(\bar{x}_q - b_x^q, \bar{y}_q - b_y^q), \\ \boldsymbol{\zeta}_q &= f_{\mathrm{sp}}([\hat{\mathbf{b}}_q \| (\bar{x}_q, \bar{y}_q) \| \mathbf{d}_q^{\mathrm{bm}}]). \end{aligned} \tag{10}$$

Compact token and readout. At decoder layer ℓ, current query feature $\mathbf{h}_q^{(\ell)}$, quality-gated roof evidence, and spatial evidence generate a fresh token:

$$\begin{aligned} \mathbf{x}_q^{(\ell)} &= [\mathbf{h}_q^{(\ell)} \| s_q \mathbf{f}_q \| \boldsymbol{\zeta}_q], \\ \mathbf{o}_q^{(\ell)} &= f_{\mathrm{tok}}(\mathbf{x}_q^{(\ell)}) \in \mathbb{R}^d, \\ \hat{\boldsymbol{\delta}}_q^{(\ell)} &= f_{\mathrm{out}}(\mathbf{o}_q^{(\ell)}) \in \mathbb{R}^2. \end{aligned} \tag{11}$$

We set $d = 7$ by default and replace only this terminal width in the sweep. The f_{feat}, f_{sp}, f_{tok}, and f_{out} contain the four width-dependent projections counted in Section 4.4. The independent map $f_{\mathrm{out}} : \mathbb{R}^d \to \mathbb{R}^2$ keeps token capacity distinct from the fixed RFO output dimension. OffsetHead is the evidence construction for compact tokenization, rather than a set of separately claimed modules.

3.4 Target-Indexed Layer-Wise RFO Supervision

The denoising motivation is to train layer-wise RFO estimates from perturbed queries while keeping each instance's clean RFO target fixed. Each ground-truth instance is replicated into denoising groups. Perturbed labels and reference boxes produce queries isolated from matching queries by an attention mask, while their construction-time target indices remain fixed. These indices align every supervised layer with the same clean signed RFO without another Hungarian assignment. Hence the successive estimates are target aligned, but they are not a recurrent update sequence and the formulation does not assume monotonic error reduction across layers. For the recorded initial output only, a fixed column-orthonormal map lifts a valid 2-D RFO to seven dimensions before additive or multiplicative perturbation; this value is not propagated, and later layers estimate RFOs from their current evidence.

For binary validity indicator v_n, matched pairs $\mathcal{M}^{(\ell)}$, known target-indexed denoising pairs $\mathcal{M}_{\mathrm{DN}}^{(\ell)}$, supervised layer set $\mathcal{S}_{\mathrm{DN}}$, denoising-group count s_{DN}, and layer weight ω_ℓ, the layer-wise offset objectives are

$$\begin{aligned} e_\delta^{(\ell)}(q, n) &= v_n \sum_{k \in \{x,y\}} \mathrm{SmoothL1}_{\beta=1}(\hat{\delta}_{q,k}^{(\ell)} - \delta_{n,k}), \\ \mathcal{L}_\delta^{(\ell)} &= \frac{1}{N_{\mathrm{gt}}} \sum_{(q,n) \in \mathcal{M}^{(\ell)}} e_\delta^{(\ell)}(q, n), \\ \mathcal{L}_{\delta,\mathrm{DN}} &= \sum_{\ell \in \mathcal{S}_{\mathrm{DN}}} \frac{\omega_\ell}{N_{\mathrm{gt}} s_{\mathrm{DN}}} \sum_{(q,n) \in \mathcal{M}_{\mathrm{DN}}^{(\ell)}} e_\delta^{(\ell)}(q, n). \end{aligned} \tag{12}$$

Here $v_n = 1$ only when target n has a valid offset annotation; $\mathcal{M}^{(\ell)}$ and $\mathcal{M}_{\mathrm{DN}}^{(\ell)}$ are respectively the Hungarian-matched and known denoising pairs. The set $\mathcal{S}_{\mathrm{DN}}$ selects supervised denoising outputs, while s_{DN} and ω_ℓ denote the number of denoising groups and the layer weight. Hungarian assignment augments the standard classification, point-sampled mask BCE/Dice, box L_1, and GIoU costs with the valid two-dimensional offset distance:

$$\begin{aligned}\mathcal{C}_{\mathrm{match}} = \lambda^m_{\mathrm{cls}}\mathcal{C}_{\mathrm{cls}} + \lambda^m_{\mathrm{mask}}\mathcal{C}_{\mathrm{mask}} + \lambda^m_{\mathrm{dice}}\mathcal{C}_{\mathrm{dice}} \\ + \lambda^m_{\mathrm{box}}\mathcal{C}_{\mathrm{box}} + \lambda^m_{\mathrm{box}}\lambda^m_\delta\mathcal{C}_\delta + \lambda^m_{\mathrm{giou}}\mathcal{C}_{\mathrm{giou}}.\end{aligned} \tag{13}$$

The offset matching term $\mathcal{C}_\delta$ is the two-dimensional L_1 distance, evaluated only for valid RFO annotations. It associates roof predictions with targets using displacement as well as class, mask, and box evidence. Matched outputs use

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

The final decoder output, auxiliary decoder outputs, and two-stage encoder candidates receive the corresponding set loss; denoising outputs additionally use $\mathcal{L}_{\delta,\mathrm{DN}}$.

Offsets are stored in original-image pixels and transformed as vectors. If geometric augmentation T produces width W_a and height H_a, then

$$\boldsymbol{\delta}^{\mathrm{aug}} = T(\boldsymbol{\delta}^{\mathrm{px}}) - T(\mathbf{0}), \qquad \boldsymbol{\delta} = \alpha\left(\frac{\delta_x^{\mathrm{aug}}}{W_a}, \frac{\delta_y^{\mathrm{aug}}}{H_a}\right). \tag{15}$$

Subtracting the transformed origin cancels translation while retaining scaling, reflection, and rotation; α controls the normalized scale. Inverse normalization and geometric transformation recover the pixel-space RFO at inference.

4 Experiments

4.1 Experimental Settings

Datasets. We use five named data sources. BONAI (Wang et al. 2022) supports the main comparison, IRSA (Meng et al. 2025) the ablations, and BONAI, IRSA, WHU (Ji, Wei, and Lu 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 $d = 7$. We optimize with AdamW (Loshchilov and Hutter 2017) for 35 epochs on six NVIDIA RTX 3090 GPUs. Inputs use a 768×768 canvas with scale jittering, flipping, and rotation. Configurations within each ablation share the same data split and training protocol.

Evaluation Metrics. Following the roof-mask/RFO protocol (Wang et al. 2022; 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.

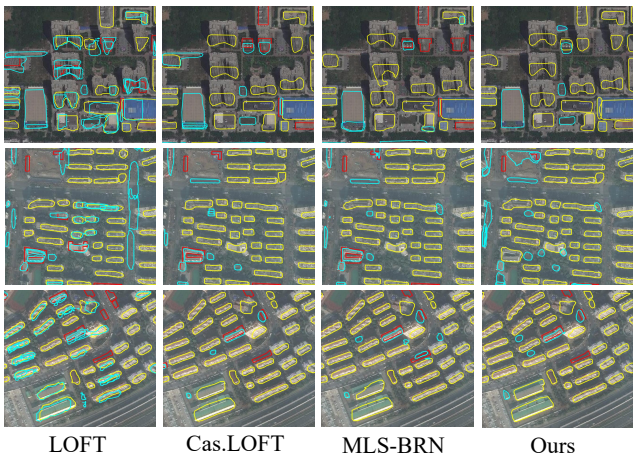

Figure 2: Qualitative comparison of LOFT, Cas.LOFT, MLS-BRN, and LoDEOT (ours) on three representative BONAI scenes. Each row shows the same scene across methods, and each column corresponds to one method.

4.2 Comparison with Existing Methods

Table 1 compares LoDEOT with two externally guided methods and three existing end-to-end baselines on BONAI. OBM (Li et al. 2024a) and PolyFootNet (Li et al. 2025) use masks predicted by LoDEOT as instance-level guidance and are included only as supplementary references. In contrast, LOFT (Wang et al. 2022), CAS-LOFT (Li et al. 2024a), and MLS-BRN (Li et al. 2024b) are automatic end-to-end methods. These are previously reported baselines with directly comparable roof-mask and signed-RFO outputs. Because the systems also differ in architecture and training, the table supports a task-level comparison rather than attributing every gain to the compact prediction terminal. Figure 2 complements the numerical results with representative visual comparisons among the end-to-end methods.

Among the end-to-end methods, LoDEOT achieves the best bAP and bAP50 scores. Compared with LOFT, CAS-LOFT, and MLS-BRN, it improves bAP/bAP50 by 18.51/12.29, 13.21/3.74, and 13.56/3.56 points, respectively. Roof segmentation is the trade-off: CAS-LOFT obtains the highest sAP and sAP50, while LoDEOT's corresponding scores are 5.99 and 9.97 points lower. For offset-corrected footprints, LoDEOT obtains 49.91 FAP50, 56.43 F1, 11.19 B-IoU, a 5.24-pixel mEPE, and 64.06% PCP@5, leading all five metrics. Because the guidance for OBM and PolyFootNet is derived from masks produced by LoDEOT, these two externally guided methods are included only as supplementary references. Nevertheless, LoDEOT outperforms both methods on all their reported metrics. The joint pattern indicates stronger roof localization and RFO recovery, but not uniformly better roof masks.

4.3 Cumulative Multi-Source Training and Cross-Domain Evaluation

We construct cumulative levels BONAI (L1), BONAI+IRSA (L2), BONAI+IRSA+WHU (L3), and

Method	Roof Detection and Segmentation				Offset-Corrected Footprint				
	bAP↑	bAP50↑	sAP↑	sAP50↑	FAP50↑	F1↑	B-IoU↑	mEPE↓	PCP@5↑
Externally Guided Methods									
OBM (Li et al. 2024a)	–	–	12.59	36.71	22.21	41.72	9.25	5.72	53.68
PolyFootNet (Li et al. 2025)	–	–	8.91	33.18	18.76	39.75	7.32	6.54	44.71
End-to-End Methods									
LOFT (Wang et al. 2022)	10.50	42.90	21.00	59.50	37.73	49.64	8.78	5.65	57.74
CAS-LOFT (Li et al. 2024a)	15.80	51.45	**30.21**	**63.25**	47.02	55.60	11.12	5.52	60.76
MLS-BRN (Li et al. 2024b)	15.45	51.63	30.20	62.81	42.73	53.99	11.17	5.67	61.02
LoDEOT(Ours)	**29.01**	**55.19**	24.22	53.28	**49.91**	**56.43**	**11.19**	**5.24**	**64.06**

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. The arrows indicate whether a higher or lower value is preferable.

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.

Each added source gives its largest immediate FAP50 gain on its own domain: IRSA rises from 8.77 at L1 to 48.76 at L2, WHU from 26.60 at L2 to 96.08 at L3, and UAV from 22.65 at L3 to 41.75 at L4. The trajectory is not monotonic across domains: IRSA and WHU peak at L2/L3, whereas BONAI, UAV, and unseen Huizhou peak at L4. L4 gives the best macro-average, 56.83 sAP50 and 53.64 FAP50, improving L1 by 20.86/20.22 points; Huizhou improves by 3.41/1.11 points. Source composition, sample count, and training exposure change together, so this is descriptive cumulative-source evidence rather than a controlled scaling law.

For a descriptive diagnostic, we apply SVD to final offset latents of commonly matched instances. The first two components explain at least 99.9678% of variance for every domain–level pair; mean cross-domain top-two principal angles for L1–L4 are 0.67°, 0.79°, 1.16°, and 1.31°. Thus the learned coordinates share a similar dominant linear subspace while cumulative training redistributes energy therein. This concentration does not make $d = 2$ sufficient: SVD measures variance after nonlinear evidence mapping, whereas the width sweep tests predictive/optimization capacity with substantially lower FAP50/F1. We do not interpret these activations as physical channels or identifiable latent semantics.

4.4 Ablation Studies

All ablations use the same IRSA split and protocol. Each width is evaluated with one seed, so differences are sensitivity results for this setting rather than significance estimates.

Offset-Token Width. Table 3 tests 17 widths: $d \in [2, 16] \cup \{32, 64\}$. Within $d \in [2, 12] \cup \{7\}$ leads all six metrics. Over the full sweep, $d = 15$ improves bAP50/sAP/sAP50 by 0.39/0.15/0.43 points, ties FAP50, and reduces bAP/F1 by 0.17/0.54. Widths 32 and 64 add no footprint gain. The four width-dependent terminal projections contain 2,734 parameters at $d = 7$ versus 5,854 at $d = 15$; thus $d = 7$ uses 46.7% of those parameters while tying the best FAP50 and giving the best F1. This supports the geometry-inspired candidate under seed 42, not a unique, minimum, or universal neural width. Counts exclude whole-model FLOPs, memory, latency, and throughput.

Layer-Wise RFO Denoising. Table 4 compares disabled supervision with multiplicative and additive initial-offset perturbations. Both enabled configurations improve every metric; additive raises bAP, sAP, sAP50, FAP50, and F1 by 2.86, 2.64, 3.49, 3.90, and 3.07 points and exceeds multiplicative by 0.38 FAP50 and 0.67 F1. This supports the tested target-indexed layer-wise configuration; additive is the default. Because "disabled" removes the supervision path as well as its perturbation, the table does not isolate each construction choice as an independent component.

4.5 Discussion

What the evidence supports. The experiments address three distinct questions. First, Table 1 compares the complete LoDEOT system with multiple methods that share the same roof-mask/RFO interface. This task-level comparison cannot by itself attribute the observed performance gains to Offset-Head. Second, Table 3 isolates the width of the four terminal projections while holding the ablation protocol fixed; it directly supports compact width selection in that setting. Third, Table 4 evaluates the complete target-indexed denoising path and its initial perturbation choices. It supports the enabled design, but the disabled row simultaneously removes both the perturbation and its associated supervision. We therefore treat mask pooling, quality gating, and box–mask geometry as the evidence pathway that makes compact tokenization possible, not as three independently validated contributions.

Interpreting dimensionality. The physical derivation and the neural results play complementary roles. The former shows that seven variables suffice to parameterize the signed physical RFO map under the stated assumptions and therefore supplies a candidate terminal width. The latter tests that candidate empirically: $d = 7$ gives the best F1 and ties the best FAP50 among 17 settings, with fewer width-dependent parameters than $d = 15$. The small advantage of $d = 15$ on three roof metrics also rules out a claim of universal optimality. Likewise, the near-two-dimensional SVD concentration describes the variance of learned tokens after nonlinear feature construction; it neither gives the token coordinates physical meanings nor replaces the predictive width sweep.

Level	BONAI		IRSA		WHU		UAV		Huizhou		Average	
	sAP50	FAP50	sAP50	FAP50	sAP50	FAP50	sAP50	FAP50	sAP50	FAP50	sAP50	FAP50
L1	53.28	49.91	9.70	8.77	49.50	48.75	31.64	28.49	35.72	31.17	35.97	33.42
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**

Table 2: Evaluation of LoDEOT with $d = 7$ under cumulative multi-source training. sAP50 and FAP50 are percentages. The average is the unweighted mean over the five test domains, and bold values indicate the best result in each column.

d	bAP↑	bAP50↑	sAP↑	sAP50↑	FAP50↑	F1↑
2	27.97	46.99	26.97	46.80	43.52	50.01
3	29.77	48.85	28.77	48.76	45.94	52.01
4	29.30	48.28	28.41	48.33	45.67	51.55
5	30.27	49.57	29.17	49.65	46.79	52.54
6	29.10	49.05	28.36	48.95	45.47	51.17
7	**30.96**	50.82	29.88	50.77	**47.64**	**53.50**
8	29.28	49.68	28.46	49.63	45.95	52.14
9	29.23	48.75	28.21	48.76	45.32	51.05
10	28.31	47.34	27.48	47.37	44.28	50.23
11	29.83	49.85	29.07	49.73	46.59	52.01
12	29.82	49.84	29.04	49.80	46.63	52.23
13	29.44	48.98	28.52	48.81	45.92	51.50
14	30.08	50.48	29.22	50.43	47.24	52.88
15	30.79	**51.21**	**30.03**	**51.20**	47.64	52.96
16	30.24	49.48	29.17	49.50	46.70	52.60
32	29.44	49.07	28.66	49.17	45.79	51.56
64	30.18	49.45	29.01	49.43	46.71	52.35

Table 3: Offset-token width ablation on the IRSA test set using random seed 42. All metrics are percentages, and bold values indicate the best result in each column.

Configuration	bAP↑	sAP↑	sAP50↑	FAP50↑	F1↑
Disabled	28.10	27.24	47.28	43.74	50.43
Multiplicative	30.85	29.66	50.33	47.26	52.83
Additive	**30.96**	**29.88**	**50.77**	**47.64**	**53.50**

Table 4: Ablation of layer-wise denoising supervision for RFO estimation on the IRSA test set with $d = 7$. Multiplicative and additive denote perturbations applied when constructing the initial offset prediction. All metrics are percentages, and bold values indicate the best result per column.

These distinctions keep the geometric argument as a width-selection prior rather than an asserted equivalence between camera variables and neural channels.

Scope and limitations. The cumulative-source study is descriptive: L4 attains the best five-domain macro-average and improves Huizhou over L1, but source composition and training-set size vary jointly, and several domains peak earlier. The results therefore support only the aggregate benefit of the evaluated mixtures, not an isolated effect of data volume. The single-seed, 17-setting width sweep supports $d = 7$ as a favorable performance–terminal-complexity choice under this protocol, not a statistically unique optimum. Because the offset terminal uses roof-mask features and box–mask geometry, uncertain masks may weaken both pooled appearance and box–mask displacement evidence; Table 1 further shows that the strongest footprint results do not coincide with the best roof-segmentation scores. Finally, the geometric motivation assumes local pinhole projection and vertical extrusion; robustness to departures from these assumptions was not evaluated.

Operational interpretation. LoDEOT does not estimate camera parameters, building height, or the seven physical variables at inference. They motivate the tested terminal width only; the predictor consumes learned image, query, mask, and box evidence and directly outputs a signed two-coordinate RFO. Denoising groups, fixed target indices, and the lifted initial offset are also training-only, so ordinary inference uses the matching-query branch and its per-layer OffsetHead predictions. The parameter comparison has the same deliberately narrow scope: 2,734 versus 5,854 counts the four projections whose shapes change with d, but does not imply a measured whole-model speed, memory, or energy gain. At the task level, a single RFO translates each predicted roof mask, which matches the annotation interface and the evaluated baselines but cannot express spatially varying displacement within one instance. Together, these points define the intended use of LoDEOT as compact instance-level offset estimation rather than camera recovery, dense geometric reconstruction, or a general compression method.

5 Conclusion

LoDEOT preserves high-dimensional multitask queries while decoupling signed RFO prediction into a compact, mask-guided terminal. Under local pinhole and vertical-extrusion assumptions, a seven-variable physical factorization motivates a candidate width while leaving learned channels unconstrained; retained denoising-query target indices supervise estimates against the same clean instance offset. The BONAI comparison, 17-setting width sweep, and denoising ablation support accurate footprint recovery and a favorable terminal performance–complexity choice; the largest cumulative source set gives the best five-domain macro-average. These findings do not establish physical meanings for learned channels, width minimality or universal optimality, whole-model efficiency, or a causal multi-source scaling law. The width study uses one seed, source composition and exposure change together, mask quality affects offset evidence, and roof segmentation remains a trade-off. Future work should test calibrated and spatially varying cameras, nonvertical structures, repeated-seed width selection, and stronger coordination between roof masks and offsets.

References

Acuna, D.; Ling, H.; Kar, A.; and Fidler, S. 2018. Efficient Interactive Annotation of Segmentation Datasets With Polygon-RNN++. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 859–868.

Adimoolam, Y. K.; Poullis, C.; and Averkiou, M. 2025. 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*, 8473–8482.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *Computer Vision – ECCV 2020*, 213–229.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-Attention Mask Transformer for Universal Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1290–1299.

Cheng, B.; Schwing, A. G.; and Kirillov, A. 2021. Per-Pixel Classification Is Not All You Need for Semantic Segmentation. In *Advances in Neural Information Processing Systems*, volume 34, 17864–17875.

Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; and Raskar, R. 2018. DeepGlobe 2018: A Challenge to Parse the Earth Through Satellite Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 172–181.

Fang, Y.; Yang, S.; Wang, X.; Li, Y.; Fang, C.; Shan, Y.; Feng, B.; and Liu, W. 2021. Instances As Queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6910–6919.

Girard, N.; Smirnov, D.; Solomon, J.; and Tarabalka, Y. 2021. Polygonal Building Extraction by Frame Field Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5891–5900.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

Hu, Y.; Wang, Z.; Huang, Z.; and Liu, Y. 2023. PolyBuilding: Polygon Transformer for Building Extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 199: 15–27.

Ji, S.; Wei, S.; and Lu, M. 2019. 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.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L. M.; and Zhang, L. 2022. DN-DETR: Accelerate DETR Training by Introducing Query DeNoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13619–13627.

Li, F.; Zhang, H.; Xu, H.; Liu, S.; Zhang, L.; Ni, L. M.; and Shum, H.-Y. 2023. 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*, 3041–3050.

Li, K.; Deng, Y.; Chen, J.; Meng, Y.; Xi, Z.; Ma, J.; Wang, C.; Wang, M.; and Zhao, X. 2025. PolyFootNet: Extracting Polygonal Building Footprints in Off-Nadir Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–16. Art. no. 5635016.

Li, K.; Deng, Y.; Deng, L.; Xi, Z.; Wang, C.; Zhang, J.; Ji, Y.; Meng, Y.; and Zhao, X. 2026a. ObliCity: A Benchmark and Baseline for Roof-to-Ground Projection Displacement Correction. *arXiv preprint arXiv:2607.25210*.

Li, K.; Deng, Y.; Kong, Y.; Liu, D.; Chen, J.; Meng, Y.; Ma, J.; and Wang, C. 2024a. Prompt-Driven Building Footprint Extraction in Aerial Images With Offset-Building Model. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–15.

Li, W.; Yang, H.; Hu, Z.; Zheng, J.; Xia, G.-S.; and He, C. 2024b. 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*, 27728–27737.

Li, W.; Zhao, W.; Zhong, H.; He, C.; and Lin, D. 2021. Joint Semantic-Geometric Learning for Polygonal Building Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3): 1958–1965.

Li, Y.; Chen, K.; Lei, Z.; Yang, Y.; Lv, J.; and Zhou, Z. 2026b. RBF-SAM: Robust Building Footprint Extraction From Off-Nadir Remote Sensing Images via Segment Anything Model. *IEEE Transactions on Geoscience and Remote Sensing*, 64. Art. no. 5631416.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022. DAB-DETR: Dynamic Anchor Boxes Are Better Queries for DETR. In *International Conference on Learning Representations*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Maggiori, E.; Tarabalka, Y.; Charpiat, G.; and Alliez, P. 2017. Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium*, 3226–3229.

Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; and Wang, J. 2021. Conditional DETR for Fast Training Convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3651–3660.

Meng, Y.; Deng, L.; Xi, Z.; Chen, J.; Chen, J.; Yue, A.; Liu, D.; Li, K.; Wang, C.; Li, K.; Deng, Y.; and Sun, X. 2025. IRSAMap: Toward Large-Scale, High-Resolution Land Cover

Map Vectorization. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–19.

Neupane, B.; Aryal, J.; Rajabifard, A.; Aravena Pelizari, P.; and Geiß, C. 2025. Multilabel Learning With ViT for Building Footprint Extraction From Off-Nadir Aerial Images. *IEEE Geoscience and Remote Sensing Letters*, 22: 1–5. Art. no. 8000605.

Wang, J.; Meng, L.; Li, W.; Yang, W.; Yu, L.; and Xia, G.-S. 2022. Learning to Extract Building Footprints From Off-Nadir Aerial Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 1294–1301.

Wei, S.; Zhang, T.; Ji, S.; Luo, M.; and Gong, J. 2023. BuildMapper: A Fully Learnable Framework for Vectorized Building Contour Extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197: 87–104.

Xu, B.; Xu, J.; Xue, N.; and Xia, G.-S. 2023. HiSup: Accurate Polygonal Mapping of Buildings in Satellite Imagery with Hierarchical Supervision. *ISPRS Journal of Photogrammetry and Remote Sensing*, 198: 284–296.

Xu, Z.; Xu, C.; Cui, Z.; Zheng, X.; and Yang, J. 2022. CVNet: Contour Vibration Network for Building Extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1383–1391.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2023. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In *International Conference on Learning Representations*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*.

Zorzi, S.; Bazrafkan, S.; Habenschuss, S.; and Fraundorfer, F. 2022. PolyWorld: Polygonal Building Extraction With Graph Neural Networks in Satellite Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1848–1857.