IMPERIAL

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

SwinH-Fuse: A Dual-Stage Multi-Sensor Transformer Framework for Urban Building Identification and Height Estimation

Elissar Al Aawar, Sofien Resifi, Juan Felipe Mendez Espinosa, and Ibrahim Hoteit

Multiphysics, CO_2, and AI4Urban-Health Event

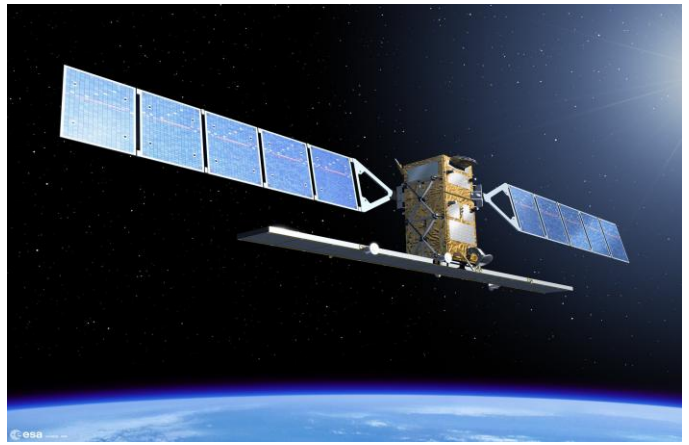
December 16th, 2025

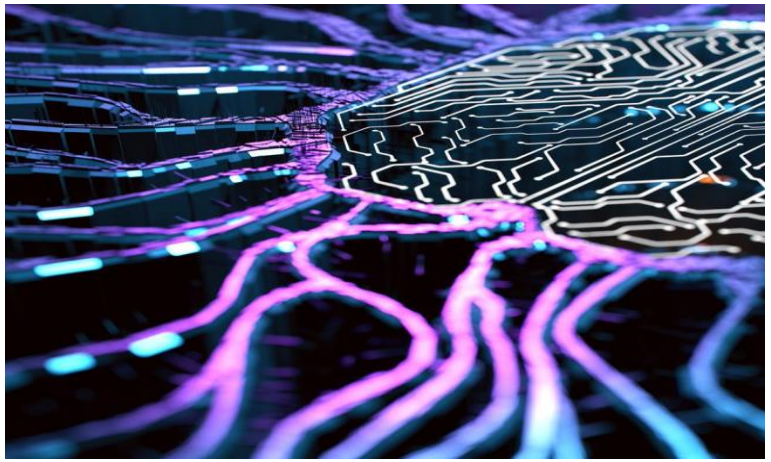
Introduction: Background

- Accurate building height information is essential for:
  - Urban morphology & 3D city models
  - Population & exposure estimation
  - Energy, microclimate, and hazard modeling
- Existing ground-truth sources:
  - LiDAR → high quality, limited availability
  - Photogrammetry → depends on coverage
  - Cadastral data → inconsistent globally
- Need: Globally scalable, automated height estimation using open-source data.

Introduction: Remote Sensing & DL Context

- Sentinel-1 SAR captures structure, geometry, roughness
- Sentinel-2 provides spectral, material, shadow information

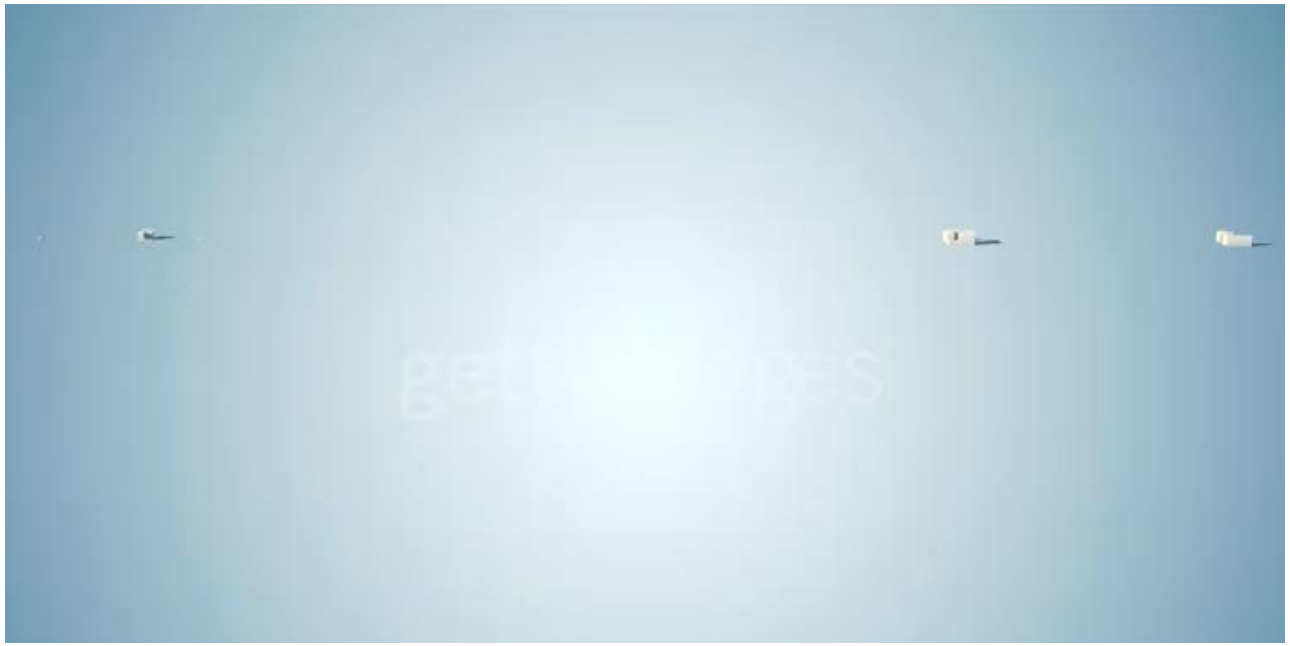




- Deep learning improves multi-sensor fusion
- Limitations of current approaches:
  - CNNs struggle with long-range dependencies
  - Limited cross-region generalization

Introduction: Problem Definition

- **Goal:** Estimate building footprints and building heights from co-registered Sentinel-1 and Sentinel-2 imagery.
- **Key challenges:**
  - Multi-sensor fusion
  - Building vs. background imbalance
  - Long-tailed height distribution
  - Generalization across diverse environments



Introduction: Contributions

SwinH-Fuse: End-to-end multi-sensor hierarchical framework

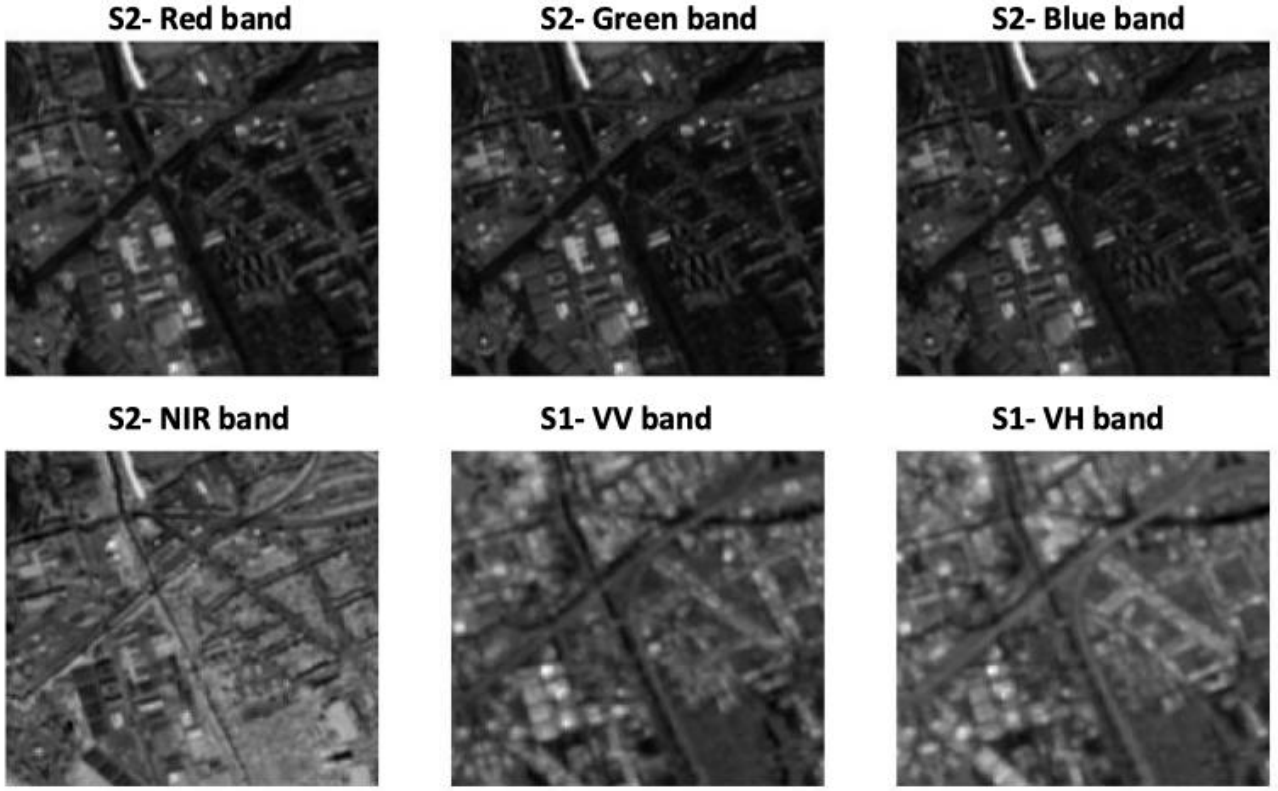
Dual-stage building classification (city-scale + neighborhood refining)

Masked log-space transformer regression

Bias correction CNN to reduce systematic height errors

Training Data: Sentinel-1/2 Inputs

- **Sentinel-1**
  - GRD IW mode
  - VV & VH polarizations
  - Multi-temporal median composites
  - Spatial resolution: 10 m
  - Sensitive to vertical structures → complements height estimation
- **Sentinel-2**
  - RGB + NIR bands
  - Level-2A surface reflectance
  - Cloud filtering & masking
  - Highlights: materials, shadows, textures
  - Improves discrimination between low-rise buildings and background



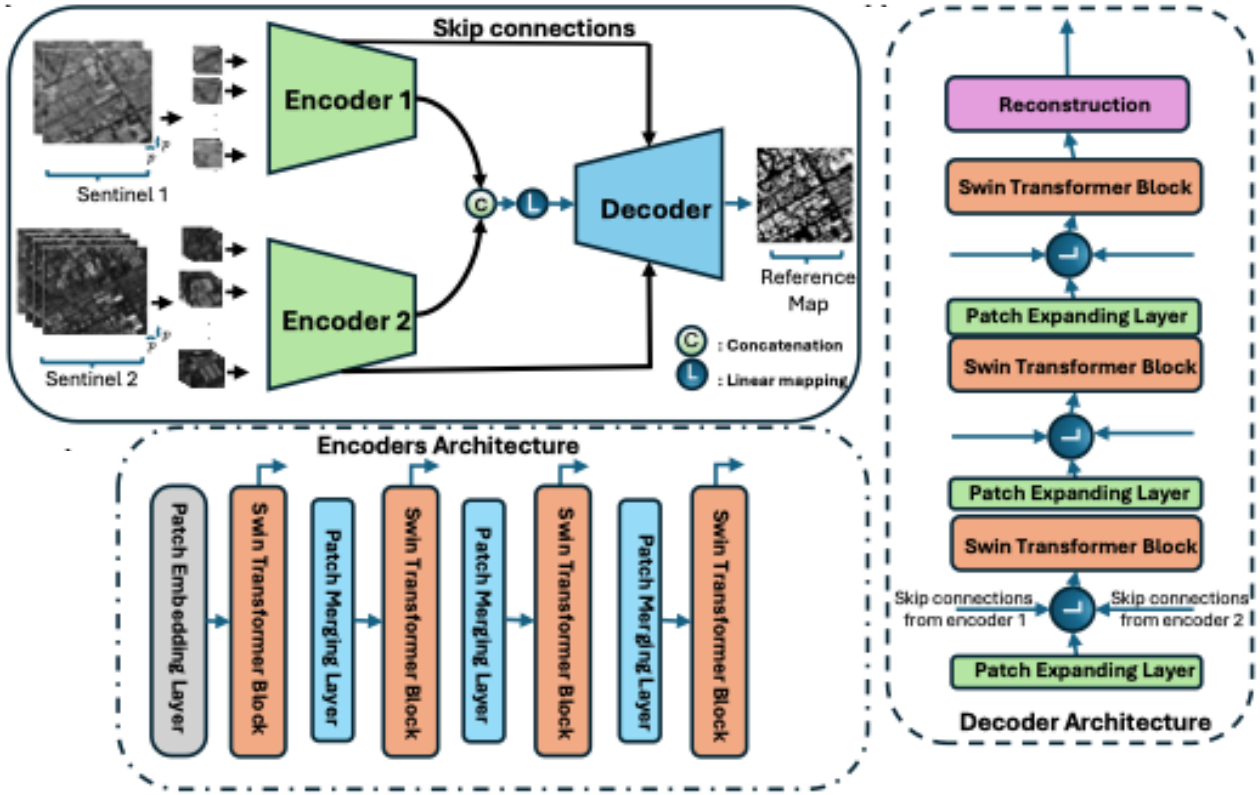
Training Data: Study Regions & Labels

- Training & validation across:
  - 51 cities in North America
  - 9 European countries
- Labels: building footprints + height rasters
- Over 2.88M patches, 128×128 pixels each

Continent	Region / Country		
North America	Arlington	Alexandria	Alberta
	Airdrie	Boston	Brookline
	Buena Park	Capetown	Chattanooga
	Chula Vista	Chester County	College Station
	Cornwall	Coromandel	Cupertino
	Dale City	Davenport	Delaware County
	Dodge County	Dublin Ohio	Evanston
	Cleveland	Fort Collins	Fairfax County
	Honolulu	Kamloops	Homestead
	LA East	LA Hub	Lethbridge
	Long Beach	Monteplier	Montgomery
	Newport News	Norman	Pasadena LA
	Peachland	Peoria	Pittsburg
	Prince George	Philadelphia	Reston
	Richardson	San Bernardino	Santa Clara
	Spokane	St. Augustine	Sunny Vale
	Tempe	West Palm Beach	Whitianga
Europe	Austria	Czech Republic	France
	Germany	Ireland	Italy
	Poland	Spain	UK

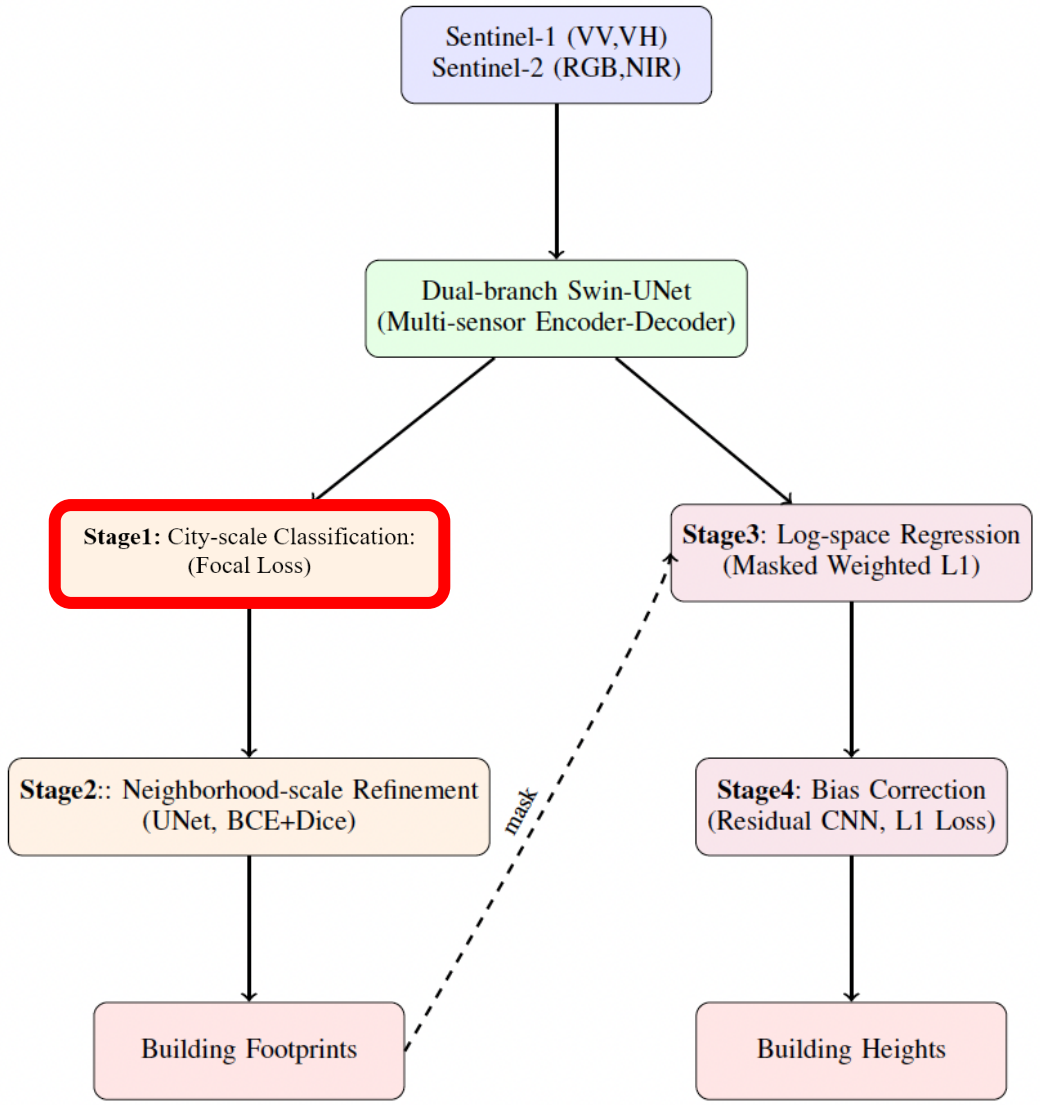
Dual-Branch Swin-UNet

- Two separate encoders:
  - SAR branch (VV, VH)
  - Optical branch (RGB, NIR)
- Shifted-window attention captures long-range spatial dependencies
- Hierarchical feature extraction
- Fused latent representation
- Decoder reconstructs spatial detail with skip connections



Framework Overview

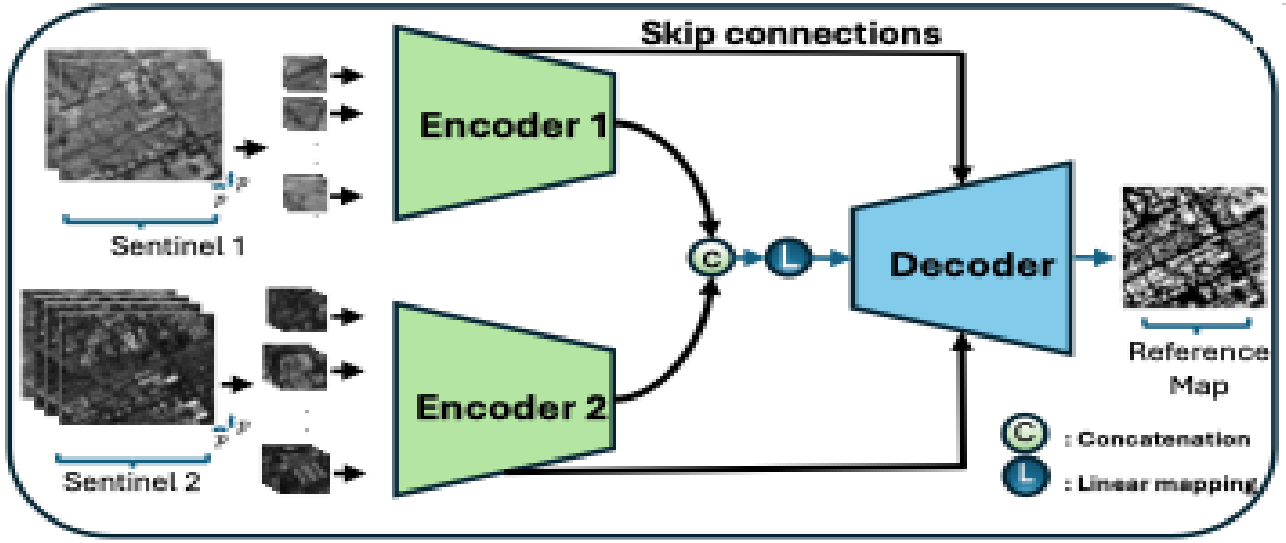
- **Four-stage architecture:**
  - City-scale classifier (Swin-UNet)
  - Neighborhood-scale refinement (Unet)
  - Masked log-space regression (Swin-UNet)
  - Bias correction CNN
- **Outputs:**
  - Building masks
  - Height maps
  - Uncertainty maps



Stage1: City-scale Classification: (Focal Loss)

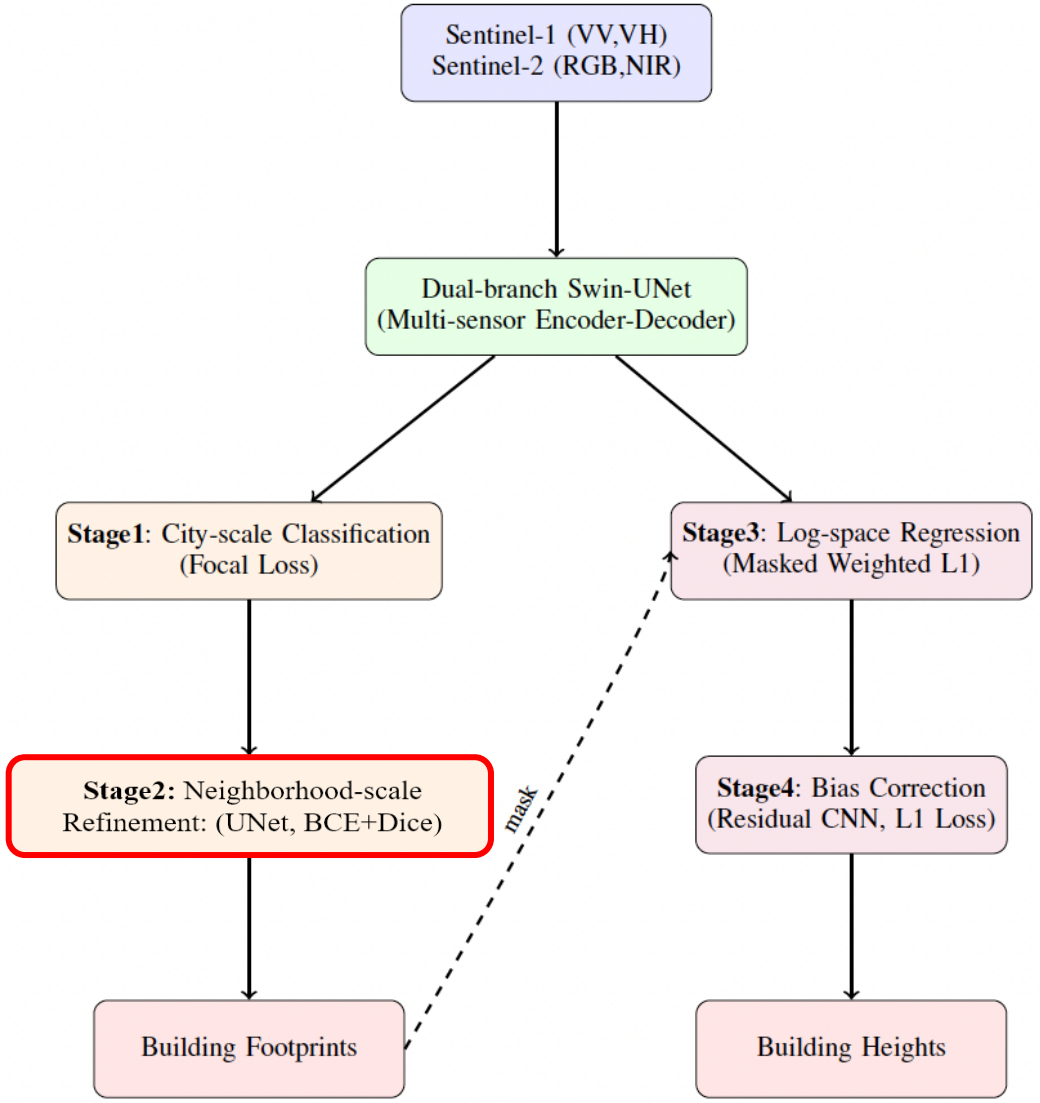
Stage 1: City-Scale Classifier

- **Purpose:** Generate coarse building map with large receptive field.
- **Design:**
  - Swin-UNet encoder–decoder
  - Focal loss for class imbalance
  - Optimized for global structure rather than fine detail
- **Output:** Initial building probability map



Framework Overview

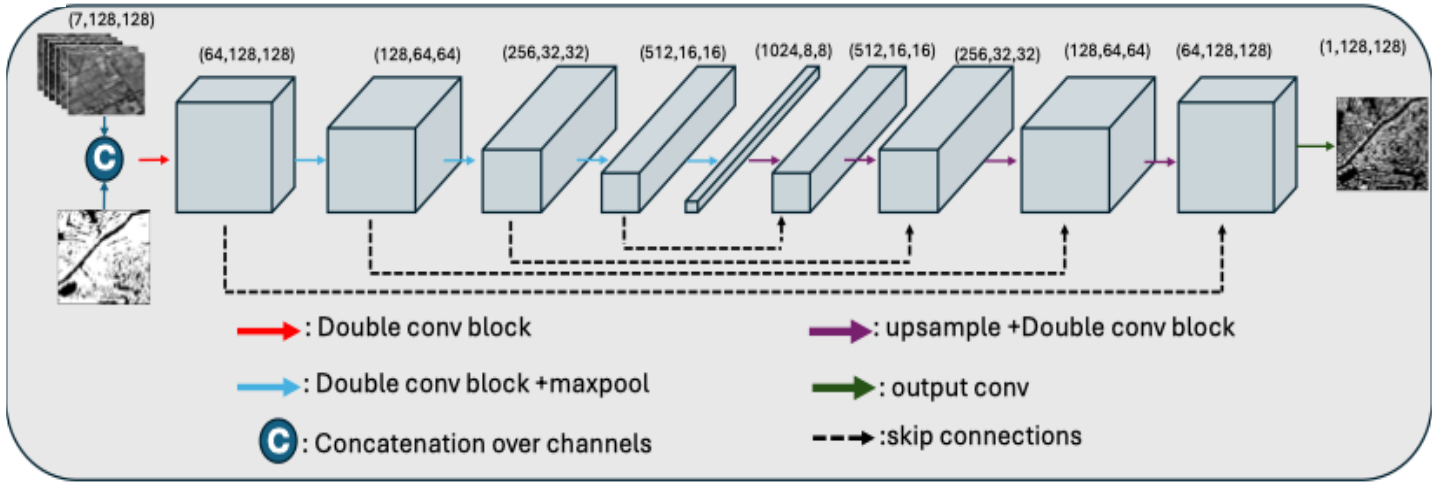
- **Four-stage architecture:**
  - City-scale classifier (Swin-UNet)
  - Neighborhood-scale refinement (Unet)
  - Masked log-space regression (Swin-UNet)
  - Bias correction CNN
- **Outputs:**
  - Building masks
  - Height maps
  - Uncertainty maps



Stage2: Neighborhood-scale Refinement: (UNet, BCE+Dice)

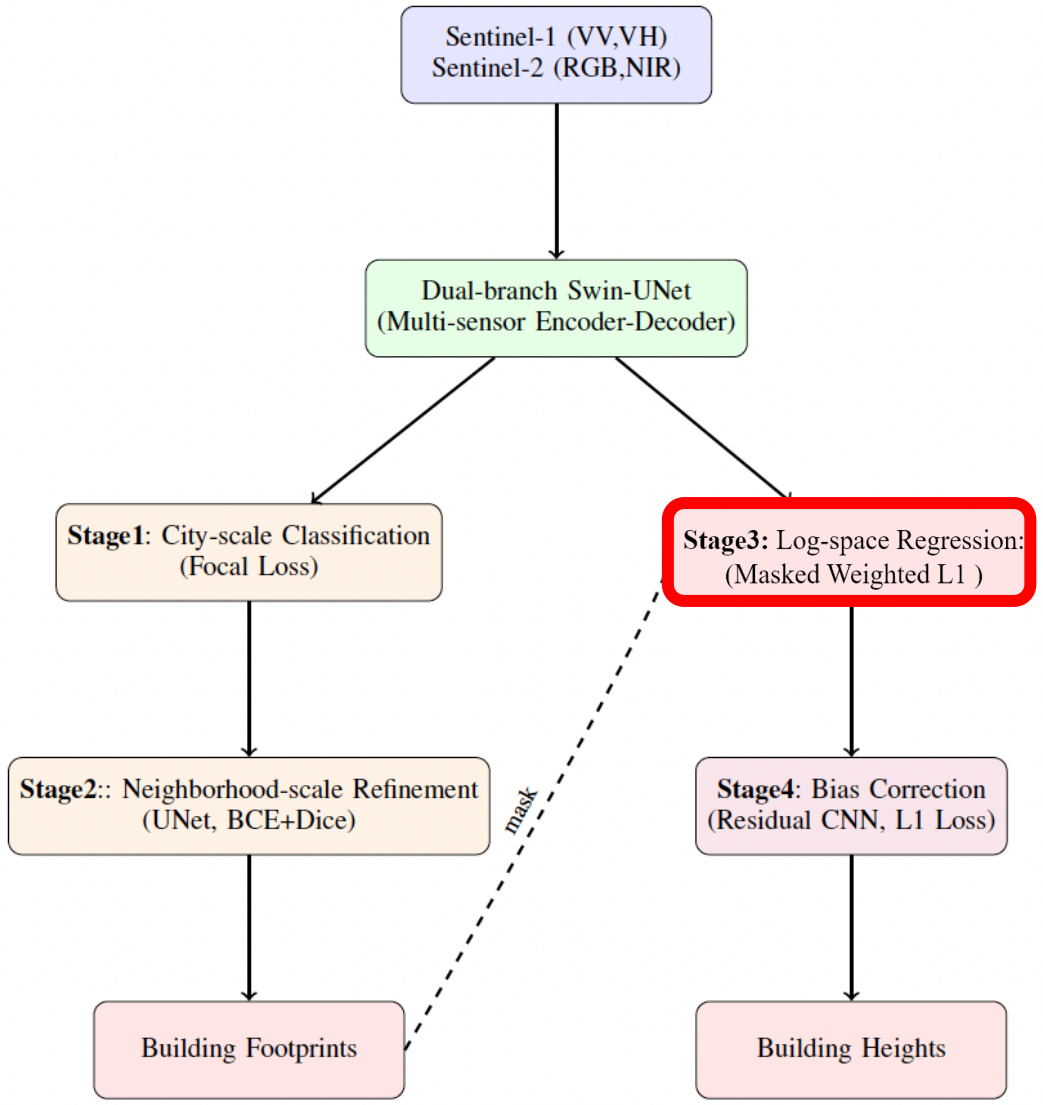
Stage 2: Neighborhood Refinement

- **Purpose:** Improve boundaries, recover small buildings.
- **Design:**
  - UNet with shallow depth
  - Loss = BCE + Dice
  - Corrects both false positives & false negatives
- **Output:** Produces final, high-quality building mask



Framework Overview

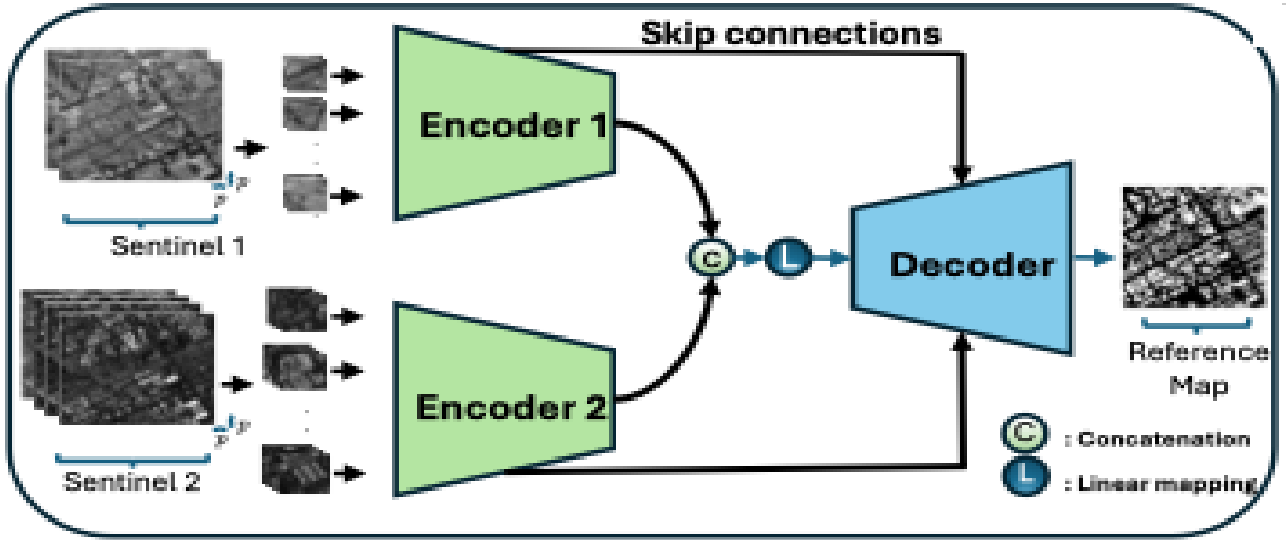
- **Four-stage architecture:**
  - City-scale classifier (Swin-UNet)
  - Neighborhood-scale refinement (Unet)
  - Masked log-space regression (Swin-UNet)
  - Bias correction CNN
- **Outputs:**
  - Building masks
  - Height maps
  - Uncertainty maps



Stage3: Log-space Regression:
(Masked Weighted L1)

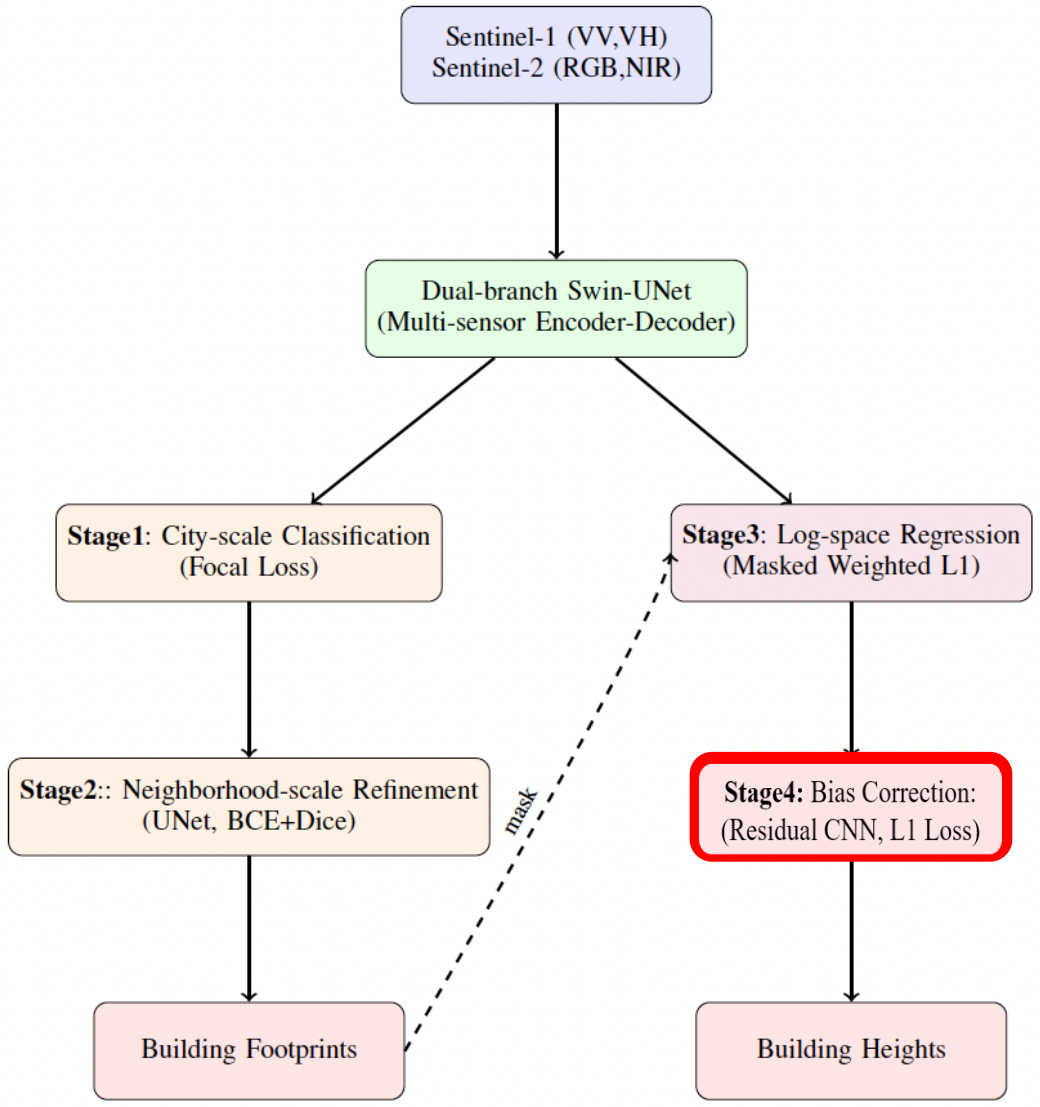
Stage 3: Regression Module

- **Objective:** estimate height on building pixels only.
- **Key components:**
  - **Log-transform:** stabilizes variance, compresses tall-building distribution
  - **Masked regression:** loss computed only on building pixels
  - **Transformer-based decoder:** same Swin backbone as classifier
  - **Weighted L1 loss:** penalizes large errors on tall buildings
- **Output:** log-height predictions



Framework Overview

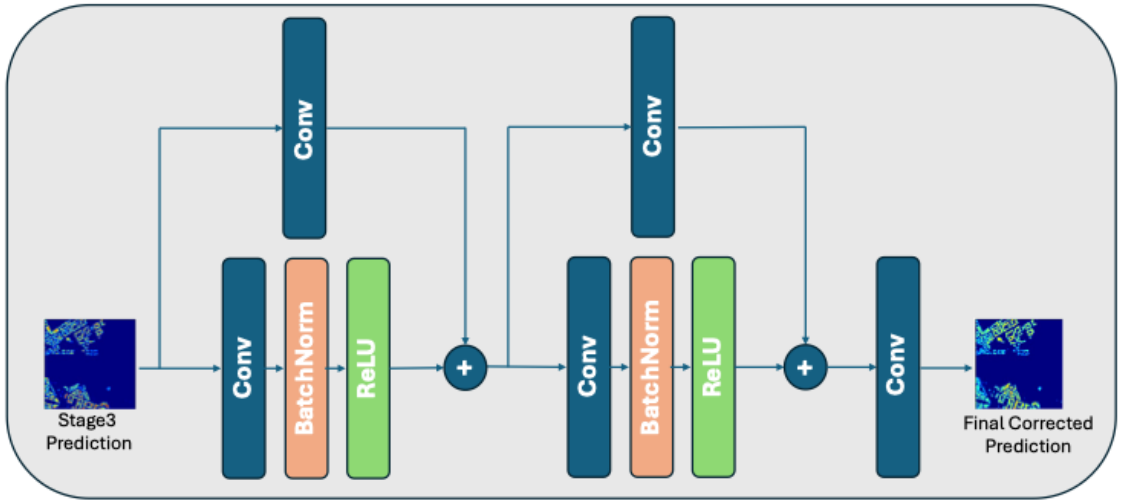
- **Four-stage architecture:**
  - City-scale classifier (Swin-UNet)
  - Neighborhood-scale refinement (Unet)
  - Masked log-space regression (Swin-UNet)
  - Bias correction CNN
- **Outputs:**
  - Building masks
  - Height maps
  - Uncertainty maps



Stage4: Bias Correction:
(Residual CNN, L1 Loss)

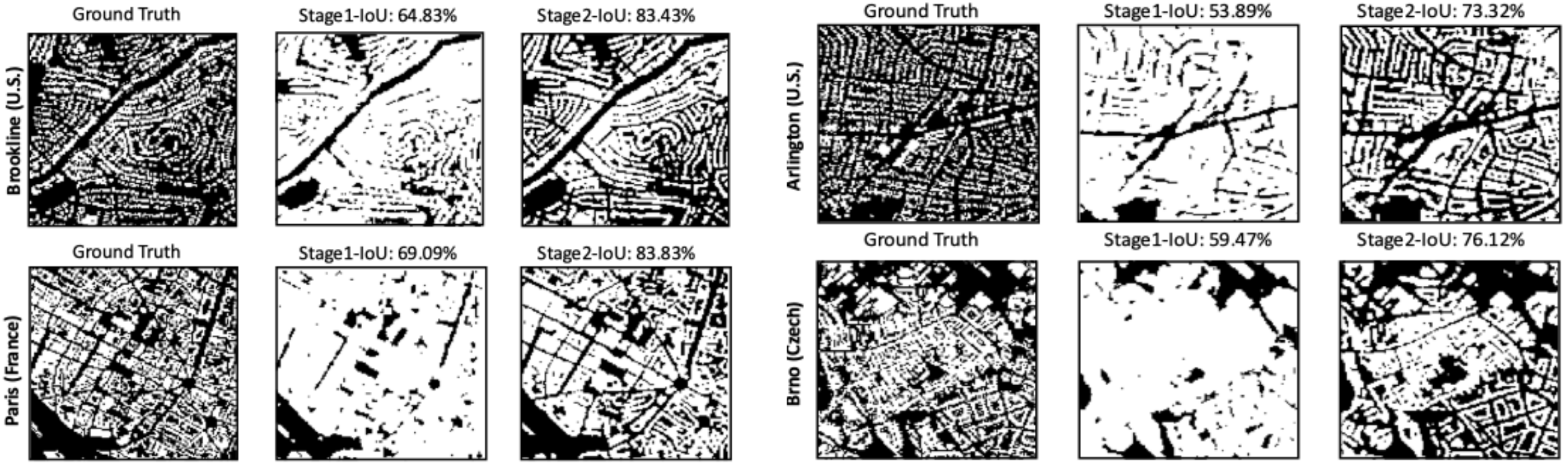
Stage 4: Bias Correction Network

- **Motivation:**
  After inverse log-transform: tall buildings often underestimated.
- **Solution:**
  - Lightweight 3-layer CNN
  - Learns residuals to correct systematic bias
  - Applied as final post-processing



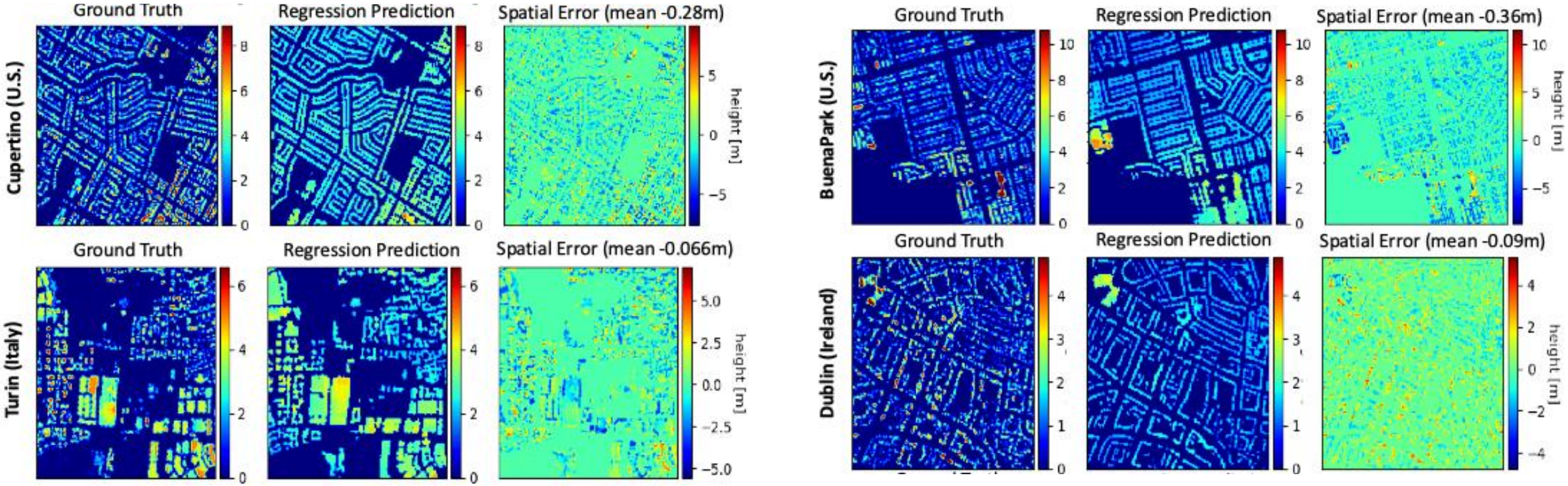
- **Output:** bias free height predictions

Building Footprint Classification Results



Metric	Stage 1 (City scale)	Stage 2 (Neighborhood scale)
IoU	0.7375	0.8064
F1-score	0.8448	0.8926
Overall Accuracy	0.8820	0.9482

Building Height Estimation Results



MBE = -0.1777 m, RMSE = 1.2199 m, and CC = 0.73

Conclusion & Future Work

- **Strengths:**
  - Multi-scale design improves both detection & height estimation
  - Combination of log-transform + masking is effective
  - Good generalization shown across continents
- **Limitations:**
  - Misclassifications propagate to regression
  - Underestimation for the tallest structures remains
- **Future Work:**
  - Higher-resolution imagery
  - Uncertainty estimation

Thank you for your time and attention!

Elissar Al Aawar
PhD Candidate, Earth Sciences and Engineering, KAUST
elissar.aawar@kaust.edu.sa