

A Semantic Segmentation Network for Urban-Scale Building Footprint Extraction Using RGB Satellite Imagery

Aatif Jiwani
 UC Berkeley
 Lawrence Berkeley Lab
 aatifjiwani@lbl.gov

Nan Zhou
 Lawrence Berkeley Lab
 nzhou@lbl.gov

Shubhrakanti Ganguly
 UC Berkeley
 Lawrence Berkeley Lab
 shubhra@lbl.gov

Chao Ding *
 Lawrence Berkeley Lab
 chaoding@lbl.gov

David Chan
 UC Berkeley
 davidchan@berkeley.edu

Abstract

Urban areas consume over two-thirds of the world’s energy and account for more than 70% of global CO₂ emissions. As stated in IPCC’s Global Warming of 1.5 °C report, achieving carbon neutrality by 2050 requires a clear understanding of urban geometry. High-quality building footprint generation from satellite images can accelerate this predictive process and empower municipal decision-making at scale. However, previous Deep Learning-based approaches face consequential issues such as scale invariance and defective footprints, partly due to ever-present class-wise imbalance. Additionally, most approaches require supplemental data such as point cloud data, building height information, and multi-band imagery - which has limited availability and are tedious to produce. In this paper, we propose a modified DeeplabV3+ module with a Dilated Res-Net backbone to generate masks of building footprints from three-channel RGB satellite imagery only. Furthermore, we introduce an F-Beta measure in our objective function to help the model account for skewed class distributions and prevent false-positive footprints. In addition to F-Beta, we incorporate an exponentially weighted boundary loss and use a cross-dataset training strategy to further increase the quality of predictions. As a result, we achieve state-of-the-art performances across three public benchmarks and demonstrate that our RGB-only method produces higher quality visual results and is agnostic to the scale, resolution, and urban density of satellite imagery.¹.

*Corresponding Author: Please send correspondance to chaoding@lbl.gov

¹Code is publicly available at <https://github.com/aatifjiwani/rbg-footprint-extract/>

1. Introduction

Urban centers consume over two-thirds of the world’s energy and account for more than 70 percent of global CO₂ emissions. Achieving carbon neutrality by 2050, as stated in IPCC’s Global Warming of 1.5°C report [1], will require a good understanding of the urban geometry at speed and scale [25]. Automatic building footprint extraction from satellite imagery is currently being pursued to support many urban science applications, such as urban planning, energy efficiency, micro-climate modeling and emergency response [10, 9, 5]. Furthermore, in high-density urban regions where buildings are often close to one another, an accurate distinction of adjacent buildings is required to provide any meaningful support to architects and urban planners. However, although building footprint extraction has received attention from the Deep Learning community [24, 30, 12, 19, 7, 34, 2], approaches based on convolutional neural networks (CNN) continue to face prominent issues such as scale invariance and defective footprints. Building density and resolution can vary significantly among satellite imagery, as can be seen in Figure 1, presenting class imbalances that make it difficult for a CNN to be robust on unseen data. Consequently, CNN-based approaches commonly suffer from sparse false positives and adjacent footprints being predicted as a single entity, drastically reducing the level of insight urban planners can glean from results.

The Deep Learning community has made many approaches that attempt to resolve conjoined and extraneous footprints, however, most approaches require supplemental data such as point clouds, building height information, and multi-band imagery - all of which are too expensive to produce or unattainable for most cities worldwide. Contrary to these methods, we propose a novel method that relies on easily attainable and globally available RGB satel-

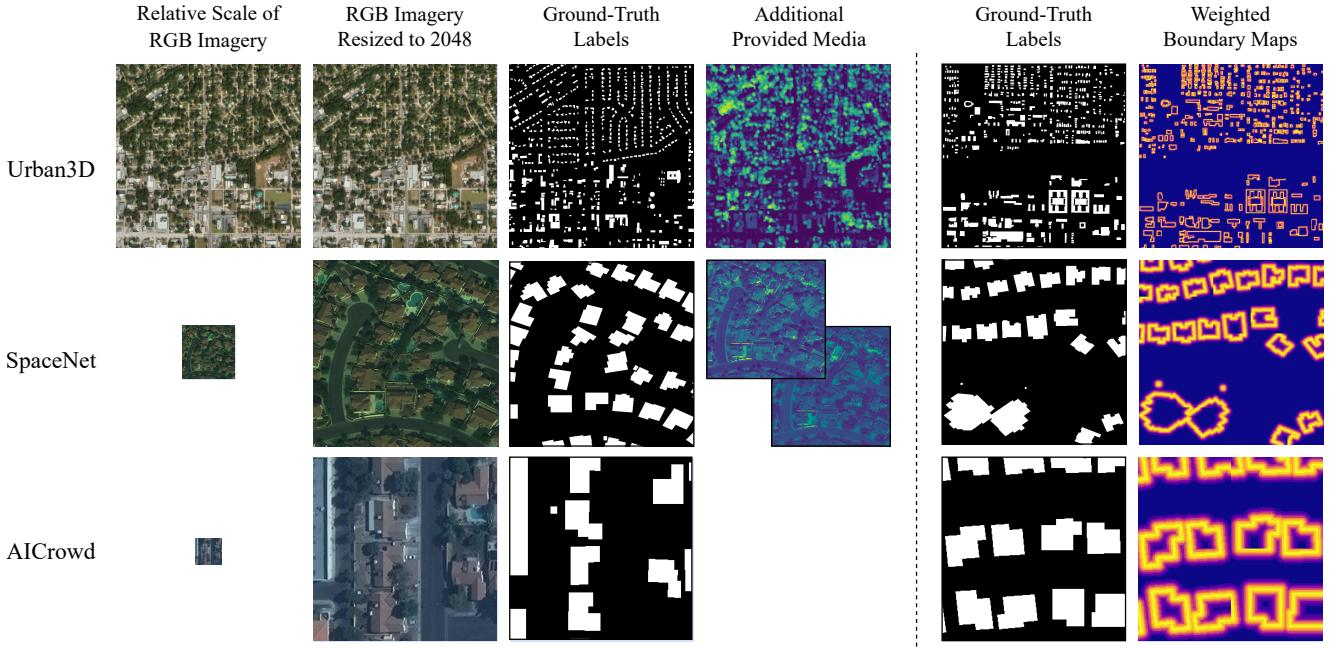


Figure 1. **Left:** Visualizations of samples from each dataset used in this paper (see Section 4.1 for details). **Right:** Visualizations of weighted boundary maps ($w_0 = 10, \sigma = 7.5$).

lite images. In the following sections, we present a modified DeepLabV3+ architecture and unique two-part training objective, and demonstrate that our method not only performs better in numerical performance, but also produces cleaner footprint predictions and precisely separates buildings in densely populated areas.

2. Related Work

2.1. Deep Learning and Building Segmentation

Building segmentation in remote sensing has long been established as an important task in urban planning, inviting robust non-Deep Learning approaches such as LIDAR reconstruction [14]. However, with processing power becoming progressively cheaper and more accessible, Deep Learning approaches for this task are becoming increasingly prominent and successful. Li *et al.* [19] tackle urban building segmentation with a U-Net ensemble and attempt to resolve the scale invariance problem by splitting their ensemble into two sets: one receives a downsampled copy, and the other receives a stack of 9 equal sections from the original image. Delassus *et al.* [7] also use a U-Net ensemble, but move in a different direction and attempt to separate footprints by employing the DICE Coefficient [29, 8]. In our work, we demonstrate that replacing the DICE Coefficient with the F-Beta measure has numerical and visual benefits. Marmanis *et al.* [23] employ a simple multi-layer perceptron, and attempt to separate conjoined footprints by fram-

ing the task as a three-class classification problem. Similarly, Yang *et al.* [31] employ a SegNet [3] and developed a signed-distance function to encode relative distances between buildings within ground-truth masks. Using the distance function, they transform the binary masks into masks with 128 classes wherein pixels that are closer to buildings have a higher class than pixels that are generally background. Inspired by these approaches, in our work we introduce an exponentially weighted boundary loss to our objective that heavily penalizes incorrect boundary predictions to achieve better separation of buildings in close proximity. Instead of manipulating the objective, Zorzi *et al.* [35] propose a pipeline that performs regularization and polygonization on segmentation masks to produce clear-cut footprints. Unlike any of the previous approaches, [35] employ generative adversarial networks with the intention of cleaning up segmented masks before polygonization [13]. However, as Zorzi *et al.* note, this method often hinders performance when buildings are occluded by the presence of greenery, creating skewed footprints that do not properly overlap with the ground-truth.

2.2. Loss Functions for Data Imbalance

Class-wise imbalance in the data is a tough problem that presents itself in multifarious tasks, but we crucially highlight its presence in object detection and image segmentation. In object detection, there is an inherent imbalance as foreground objects are often just a small fraction of the

full image. Lin *et al.* [20] address the difficulty of training object detectors with the Focal loss. Commonly applied in object detection tasks, the Focal loss builds upon binary Cross-Entropy by adding a tunable factor that effectively puts more emphasis on penalizing misclassifications by shifting focus away from the common class. However, the Focal loss presupposes the common class is unimportant, which does not translate well to building segmentation as we observed copious false positives footprints. In binary image segmentation, the common problem of class imbalance poses a downstream problem for either precision or recall. Salehi *et al.* [28] developed the Tversky index to shift the network’s focus on resolving false negatives. The Tversky index, based upon the DICE coefficient [29, 8], is an ornate function of the predicted and ground-truth masks with tunable weights α and β that shift emphasis towards precision and recall respectively. Compared to [20] and [28], we introduce the F-Beta measure which requires only a single parameter β to analogously shift focus towards precision or recall while not losing sight of the common class. Finally, Kervadec *et al.* [18] proposed a distance-function-based boundary loss to resolve imbalances in brain scans. While [18] have shown they improved the visual quality of results, the issue of separating distinct entities remains extant. In this work, we introduce a boundary loss aimed to ameliorate the separation of distinct ground-truth footprints.

3. Approach

We propose a network architecture and training method closely modeled on [19] and [7], however, with several major alterations to account for the strict use of RGB imagery:

1. We replace the U-Net encoder-decoder model with a DeepLabV3+ module [4].
2. We swap the Aligned Xception model in the DeepLabV3+ with a Dilated ResNet [32].
3. We generalize the DICE loss to an F-Beta Measure and add an exponentially weighted boundary loss to create a unique two-part training objective.
4. We perform cross-dataset training to optimize performance on multiple datasets.

We depict our overall training pipeline in Figure 2 and compare it side-by-side against the default DeepLabV3+ module.

3.1. Modified DeepLabV3+ Network

In our work, we found that for RGB data, as opposed to wideband satellite imagery, the DeepLabV3+ architecture outperformed the U-Net based method. We hypothesize that this is due to DeepLabV3+’s increased ability to model local and global relationships between parts of the visual

data, which stems from two key features in the architecture. First, the spatial pyramid pooling helps encode multi-scale contextual information in the model. Second, skip connections in the encoder, with low-level features from the backbone, help refine the prediction by allowing details (edges and corners of buildings) to permeate the final layers.

We also found that by further replacing the Xception Module in the DeepLabV3+ architecture with dilated convolutions, we can reduce the loss of spatial context information by introducing a larger receptive field (RF) for each kernel image. This added scale invariance is clear in our approach (See Figure 4). Dilated convolutions consist of kernels with “holes” in them to achieve better performance in downstream tasks, allowing for a larger RF in earlier layers without drastically increasing the number of parameters. Downsampling and pooling in the encoder can reduce spatial context in deeper layers, which the larger RF mitigates and allows us to track less prominent features (ex. smaller or narrow buildings).

3.2. Objective Function

$$\mathcal{L}(y, \hat{y}) = \mathcal{L}_{BCE}(y, \hat{y}) + (1 - D(y, \hat{y})) \quad (1)$$

Methods for semantic segmentation commonly optimize Equation (1), where D is the DICE coefficient (a generalization of the F1 Score)[17]. Works that utilize Eqn. (1) achieve relatively high numerical accuracy on the satellite segmentation datasets, but encounter visual issues such as combined footprints and sparse false building predictions (See Figure 4 ”BCE + 1” column). This hinders the analysis of urban areas because we may not obtain an accurate qualitative prediction of the building masks required for downstream energy estimation. We speculate these visual inaccuracies are due to two major causes. First, since the resolution of the data is not always high, the building-to-background distinction can be blurry, making it hard to differentiate tightly packed buildings. Second, false-positive building clusters are abundantly predicted in footprint masks, which stems from the distribution of building-to-background pixels being uneven (see Figure 1), with far more background than building. We address both of these drawbacks with a modified two-part loss formulated as follows:

$$\mathcal{L}(y, \hat{y}) = \mathcal{L}_{EWC}(y, \hat{y}) + (1 - F_\beta(y, \hat{y})) \quad (2)$$

where

$$F_\beta(y, \hat{y}) = \frac{(1 + \beta^2)(y * \hat{y})}{\beta^2(y + \hat{y})} \quad (3)$$

and

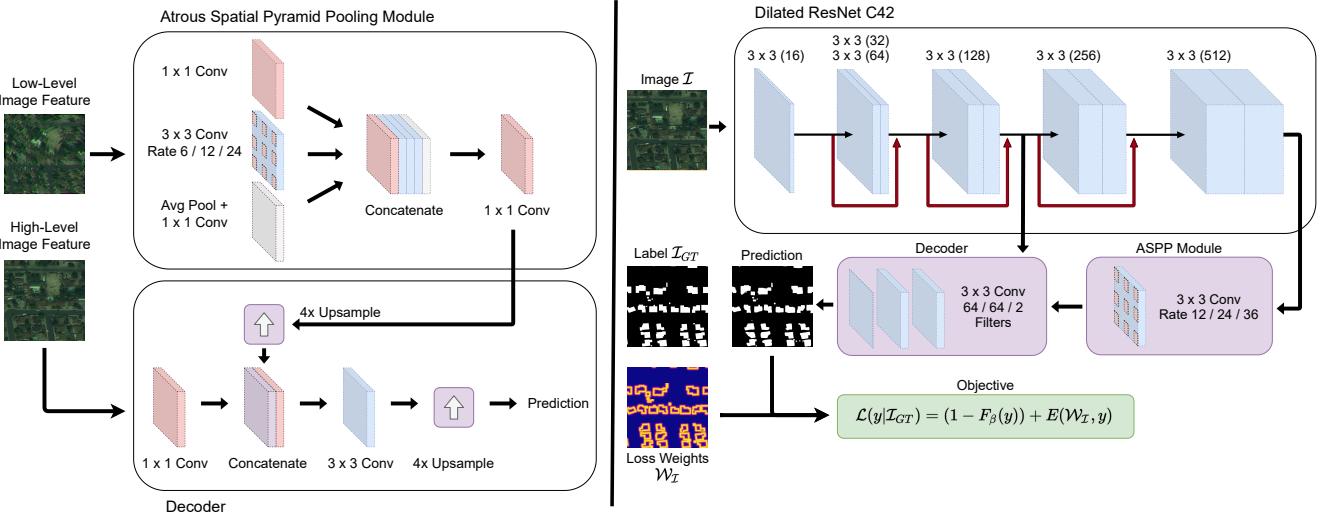


Figure 2. **Left:** DeepLabV3+ Module as depicted by [4]. **Right:** Training pipeline depicting our modified semantic segmentation network.

$$\mathcal{L}_{WCE}(y, \hat{y}) = \sum_{(i,j)} \mathcal{W}_{ij}(y) * \log(p_{y_{ij}}(\hat{y}_{ij})) \quad (4)$$

F-Beta Measure as a Loss Function: Due to imbalances in the training data, high-parameter models such as the DeepLabV3+ tend to lean toward predicting buildings rather than the more common background classes. To compensate for the erroneous false positives, we introduce the F-Beta Measure to replace the DICE coefficient in our two-part objective. Rooted from the general F1 Score, we define the F-Beta Measure in Equation (3) as a configurable loss function with a hyper-parameter β . However, although Eqn. (3) is a weighting of [8], this weighting helps the F-Beta Measure serve as a crucial **generalization** of the DICE coefficient. Note that when we set $\beta = 1$, the F-Beta Measure precisely becomes the DICE coefficient. Here, we argue that this simple yet unique formulation of the F-Beta Measure allows us to **tune** our objective toward prioritizing either false positives or false negatives. We will show that by setting $\beta > 1$, we place a higher emphasis on preventing false negatives, thus increasing recall. Likewise, if we set $\beta < 1$, we force the network to focus on improving precision, consequently resolving false positives and thus reducing the number of misplaced building clusters.

Exponential Weighted Boundary Loss (EWC): Similar to [23], which introduces a third building boundary class, we add a weighted loss function [27] that heavily penalizes wrong predictions close to and on building edges. Our exponential boundary loss generates a weight map $w(\mathbf{x})$ that forces the model to separate neighboring objects into distinct entities:

$$w(\mathbf{x}) = w_c(\mathbf{x}) + w_0 \cdot \exp\left(-\frac{(d_1(\mathbf{x}) + d_2(\mathbf{x}))^2}{2\sigma^2}\right) \quad (5)$$

where \mathbf{x} is a ground-truth mask, $w_c(\mathbf{x})$ is a separate weight map used to resolve class imbalance, and w_0 and σ are constants that help tune the absolute loss value and relative decay respectively. Using this mask not only helps us distinguish adjacent buildings, but also increases the overall loss in high-density areas (ex. Figure 1 top-right) where models traditionally exhibit a greater level of uncertainty (see Section 5).

The evaluation of Equation (5) empirically produces values between 0 and w_0 per-pixel. Although we want to heavily penalize incorrect boundary predictions, we still want to properly penalize *all* incorrect background predictions. Therefore, we build upon Equation (5) to produce a new weight map $\mathcal{W}(\mathbf{y})$

$$\mathcal{W}(\mathbf{y}) = \exp(w(\mathbf{y}))^p \quad (6)$$

that produces values between 1 and e^p , where p is an additional hyper-parameter to further prevent incorrect boundary predictions without affecting how we consider less important background predictions. Figure 1 visualizes these new weight maps on the three standard datasets. We incorporate $\mathcal{W}(\mathbf{y})$ as part of the objective in Equations (2) and (4) and claim it allows the network to separate previously conjoined predictions.

Cross-Dataset Training Strategy: As Section 4.1 will quantify, there is a stark imbalance in the number of samples across the datasets. Consequently, we saw quick over-fitting

and lack of quality results in Urban3D. To account for this lack in overall performance, we adopt a cross-dataset training strategy [33]. We first train on a synthesized set that combines augmented samples from the datasets, and then further fine-tune on the individual dataset. We claim and subsequently prove that by training on similar satellite images before fine-tuning on the individual scale, we achieve better numerical performance and more distinct visual results.

4. Experiments

4.1. Datasets

We evaluate our method on three datasets. Each dataset is split into an (80-20)-20 partition, where 20% is reserved for the final test set, and the remaining 80% is further split into (80-20) for training and validation sets, respectively.

SpaceNet: The SpaceNet.ai Building Detection Dataset (SpaceNet) [30] contains 700,000 building footprints in five cities scattered across the globe. In this study, we evaluate using the Vegas region which contains over 3,800 samples of 200m x 200m areas. Each input sample comes with an associated 650 x 650 pixel RGB satellite image, a high-resolution panchromatic image, and a low-resolution multispectral image. For our purposes, we only use the RGB satellite image. The samples in their original form come in a TIFF raster graphics format and the labels in geographical coordinates, so we use GDAL to pre-process all samples by extracting RGB rasterizations and converting the coordinates into grayscale masks [11].

AICrowd: The AICrowd Mapping Challenge dataset (AICrowd) [24] contains 340,000 total samples as 300 x 300 pixel RGB images. RGB image samples are provided in JPEG format, and annotations are provided in MS-COCO format for which we use the corresponding API to extract masks [21].

Urban3D: TopCoder’s Urban3D dataset (Urban3D) [12] contains 236 samples of 2048 x 2048 pixel images and labels, both in the TIFF compressed raster format. We use similar pre-processing steps as SpaceNet to extract RGB images and grayscale masks. As a similar note to SpaceNet, each RGB sample in this dataset is accompanied by its Depth Surface Model and Digital Terrain Model, which provides high-resolution building height information. As mentioned with SpaceNet, we use the RGB image for our purposes only.

4.2. Experimental Evaluation and Implementation

Evaluation: To evaluate each of our experiments on the datasets mentioned above, we consider the standard accuracy metrics in the fields of both semantic segmentation and binary classification: class-mean intersection-over-union (mIOU) and the F1 Score. The F1 score, also known

	Model	AP	AR	F-1	mIOU
Urban3D	Dilated-RN_D	82.0	78.5	79.6	81.0
	Res-Net _D	81.5	77.3	78.6	80.1
	U-Net	83.1	76.2	78.6	80.5
	FCN	82.7	75.5	78.2	80.0
SpaceNet	Dilated-RN_D	90.0	91.1	90.6	89.1
	Res-Net _D	90.2	89.9	90.1	89.0
	U-Net	90.6	89.1	89.7	88.2
	FCN	90.1	88.9	89.5	88.4
AICrowd	Dilated-RN_D	90.7_{.02}	90.0_{.01}	90.4_{.04}	88.6_{.00}
	Res-Net _D	90.2 _{.10}	89.0 _{.02}	89.7 _{.07}	87.9 _{.00}
	U-Net	88.4 _{.16}	86.3 _{.06}	87.0 _{.08}	85.3 _{.00}
	FCN	89.6 _{.02}	88.1 _{.00}	88.8 _{.02}	86.9 _{.00}

Table 1. Results of different network architectures (Subscript *D* indicates DeepLabV3+)

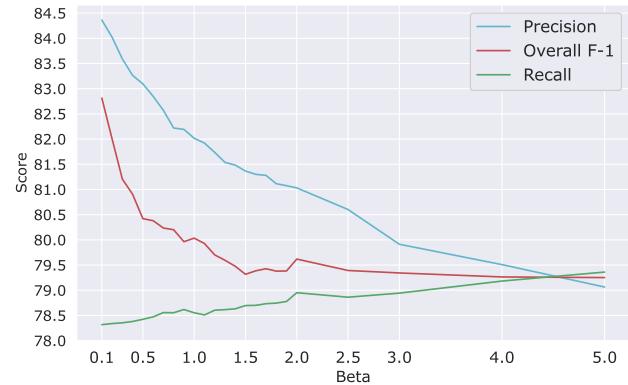


Figure 3. Avg. accuracy performance of three runs on different values of β on Urban3D

as the Sørenson-Dice coefficient [29, 8] measures the balance between precision, the ratio between true positives and all positive predictions, and recall, the ratio between true positives and all ground-truth positives. In Section 5, we also report the average precision and recall. The mIoU metric [16] is a measure of the similarity between two sets using the ratio between the intersection and union of the prediction and ground-truth.

General Training Details: The experiments concerning our proposed method are trained for 60 epochs using a batch size of 4 with standard Stochastic Gradient Descent with a learning rate of $1e^{-4}$ and an L2 constant of $1e^{-4}$. Empirically, we found that a power of 2 when using the exponentially weighted boundary masks performs best. We select the checkpoint epoch based on the best validation loss and mIoU score.

Because Urban3D inputs are too large to fit into memory, we employ a tiling operation with a factor of 4. Each input 2048 x 2048 sample image is split into 16 inputs of 512 x 512 pixels with the same image resolution. For **only the ablation studies**, we introduce a constant amount of bias in training with the AICrowd dataset by using only 15% of the

Loss	AP	AR	F-1	mIOU
U+BCE	80.09	78.04	78.88	80.31
U+F(1.0)	80.99	78.31	79.16	80.66
U+BCE+F(0.1)	84.31	77.91	82.79	83.60
U+BCE+F(1.0)	82.05	78.55	79.65	81.00
U+BCE+F(4.0)	79.50	79.02	79.25	80.64
S+BCE	90.13	90.44	90.22	88.69
S+F(1.0)	90.16	90.91	90.48	88.96
S+BCE+F(0.1)	92.16	90.37	92.04	90.65
S+BCE+F(1.0)	90.04	91.17	90.61	89.18
S+BCE+F(4.0)	89.30	91.76	90.47	88.96
C+BCE	90.19 _{.06}	88.42 _{.05}	89.18 _{.08}	87.42 _{.08}
C+F(1.0)	90.43 _{.04}	89.20 _{.05}	90.20 _{.05}	88.55 _{.03}
C+BCE+F(0.1)	91.71_{.02}	89.53 _{.03}	91.55_{.03}	89.96_{.02}
C+BCE+F(1.0)	90.74 _{.02}	90.08 _{.01}	90.40 _{.04}	88.64 _{.00}
C+BCE+F(4.0)	89.64 _{.04}	90.79_{.01}	90.25 _{.02}	88.60 _{.02}

Table 2. Ablation study of BCE and F-Beta, including both losses in isolation.

dataset. We introduce this bias both to speed up the training process, and to continue to analyze relative performance. To account for variance, we report the mean and standard deviation (reported in subscript) across 3 experiments for each study in Tables 1 - 3.

Cross-Dataset Training Details: With cross-dataset training, we augment each RGB sample to 256 x 256 pixels when combining the samples from all datasets. For Urban3D and SpaceNet, we down-sample images to 512 x 512 and then split each into 4 inputs. For AICrowd, we just resize to 256 x 256. After training on the combined set for 20 epochs, we save the model at the epoch with the best loss, and then further train on each of the three datasets individually. As an implementation note, our final method was trained using this combined set for only 5 epochs and included only 10% of the AICrowd training set.

Experiments are conducted on GPU nodes where each consists of an 8-core Intel(R) Xeon(R) CPU @ 3.00 GHz, 64 GB of RAM, and either 4 GeForce RTX 2080 Ti GPUs or 2 Tesla V100 GPUs for training models in parallel. We use PyTorch 1.6.0 to develop our methods, re-implement previously published methods, and build the training pipeline [26].

5. Results

In most tables, we use shorthand notation for datasets: U-Urban3D, S-SpaceNet, C-AICrowd.

5.1. State-of-the-Art Comparisons

We compare our method to current state-of-the-art segmentation networks by dataset. For AICrowd and Urban3D, we follow [35] which use an R2U-Net and Mask-RCNN followed by an extensive regularization and polygonization process. For only AICrowd, we follow the approach detailed by the winning submission that used a U-Net with

Loss	AP	AR	F-1	mIOU
U+EWC	76.40	82.27	78.58	79.97
U+EWC+F(0.1)	83.82	81.10	83.01	84.22
U+EWC+F(0.5)	81.49	81.14	80.75	81.87
U+EWC+F(1.0)	80.09	81.72	80.26	81.40
U+EWC+F(4.0)	78.98	81.90	79.68	80.95
S+EWC	88.90	91.51	90.13	88.55
S+EWC+F(0.1)	92.70	91.70	92.41	90.94
S+EWC+F(0.5)	90.41	91.74	91.02	89.53
S+EWC+F(1.0)	89.36	92.19	90.70	89.37
S+EWC+F(4.0)	89.25	92.34	90.52	88.98
C+EWC	89.80 _{.04}	90.05 _{.07}	89.85 _{.03}	88.00 _{.04}
C+EWC+F(0.1)	92.04_{.05}	91.06 _{.08}	91.67_{.07}	90.61_{.04}
C+EWC+F(0.5)	91.02 _{.03}	91.24 _{.10}	91.45 _{.02}	89.72 _{.01}
C+EWC+F(1.0)	89.59 _{.05}	91.63 _{.03}	90.50 _{.05}	88.90 _{.02}
C+EWC+F(4.0)	89.38 _{.02}	91.82_{.08}	90.39 _{.04}	88.83 _{.06}

Table 3. Ablation study of EWC and F-Beta, including EWC in isolation.

a ResNet-101 encoder, in total containing 51.52M parameters [6]. Then for Urban3D, we follow the winning approach which similarly takes advantage of a U-Net architecture with a ResNet-34 encoder (24.43M parameters) but requires additional depth elevation models. We report the performance of this architecture with and without the supplemental data. For SpaceNet, we follow [34], an approach similar to [35] that employs a Mask-RCNN (87.33M parameters) and a boundary regularizer in post-processing, and [7], an approach that requires two forward passes in a U-Net Ensemble (31.53M parameters). We note that [7] takes advantage of multi-spectral images which offer information through 8-channels, **while our model uses only 3-channel RGB images**. Our final network contains 58.26M parameters. Table 4 reports the results from all methods, and we also show the results of our final method with cross-dataset training. Additionally, we report the IOU on the building class ($\text{IOU}_{y=1}$) to compare with [35]. We show that despite data-level disadvantages our method performs better than all architectures, notably including the U-Net architectures across all datasets. We also note that compared to [34] and [35], we do not require any post-processing after segmentation. Additionally, while the U-Net + DEM nears our final performance, it is at a significant advantage with an extra channel (see Figure 1). The results we have presented in Table 4 quantitatively show our proposed method is state-of-the-art, requiring a single end-to-end network and no additional channels of information.

5.2. Network Architecture

Table 1 compares our base network and the three baseline architectures. We compare our DeepLabV3+ and Dilated ResNet network (DRN_D) [4, 32] against three primary baselines: the DeepLabV3 + ResNet-101 (RN_D) [15], as it is the most common pairing in DeepLabV3+ related literature, a standard U-Net encoder-decoder as it is the base

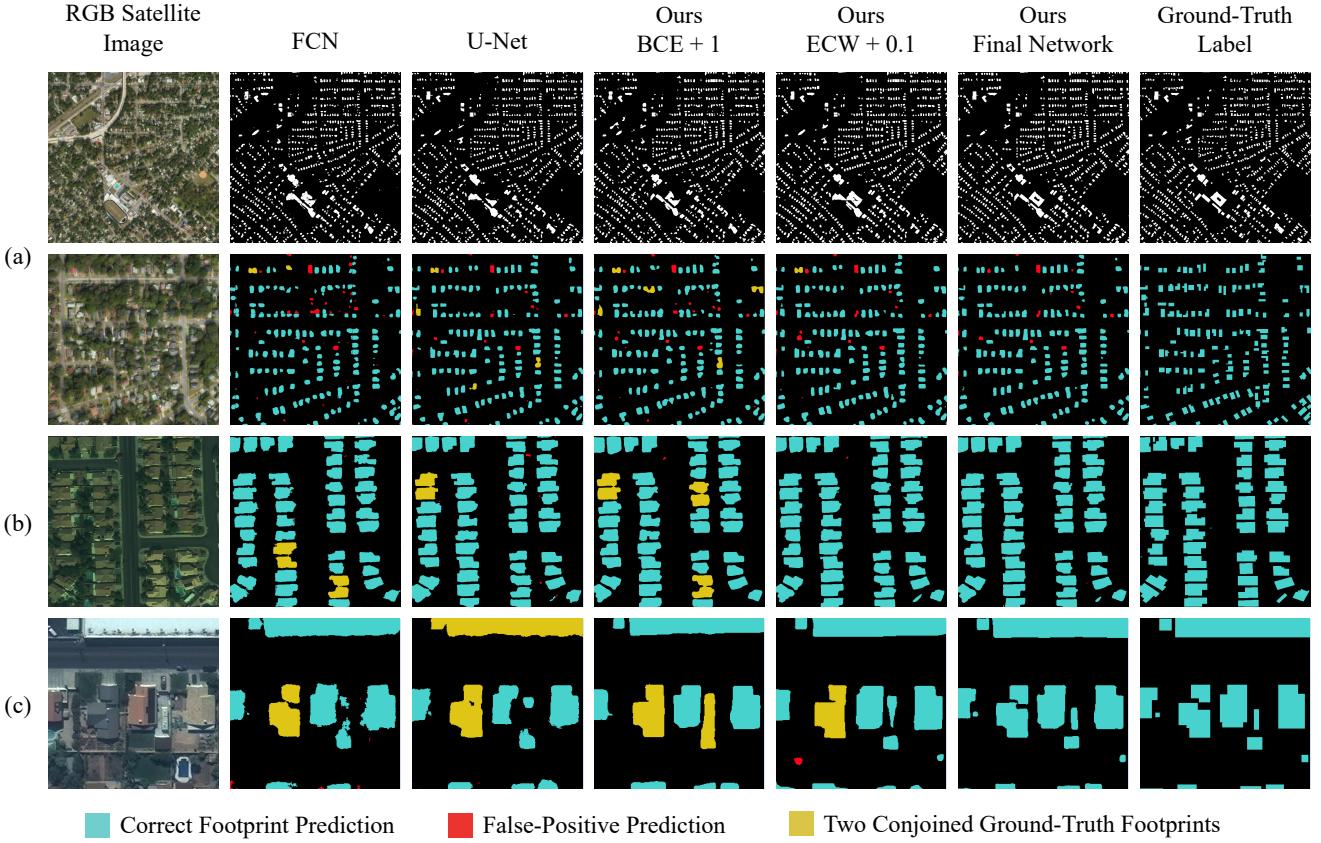


Figure 4. Visualizations of predictions from all datasets by four variants of our method and two baselines. **Top-Down:** (a) Urban3D [second row is a magnification of the first row for better visual comparison], (b) SpaceNet, (c) AICrowd. (Best viewed electronically)

Dataset	Method	AP	AR	F-1	mIOU	$\text{IOU}_{y=1}$
Urban3D	R2U-Net [35]	-	-	-	-	69.9 ± 7.1
	U-Net	84.6	79.2	81.0	80.8	65.9 ± 15.1
	U-Net + DEM	85.3	80.0	81.8	83.0	68.7 ± 15.5
	Ours	83.8	81.1	83.0	84.2	70.1 ± 15.2
	Ours + Cross-Dataset	84.5	81.3	83.4	84.5	70.2 ± 15.3
SpaceNet	U-Net Ensemble [7]	-	-	86.4	-	-
	Mask-RCNN [34]	-	-	88.1	-	-
	Ours	92.7	91.7	92.4	90.9	85.2 ± 3.5
	Ours + Cross-Dataset	93.2	92.0	92.6	91.1	85.5 ± 3.6
AICrowd	Mask-RCNN [35]	-	-	-	-	74.2 ± 14.5
	R2U-Net [35]	-	-	-	-	80.0 ± 14.2
	U-Net [6]	94.3	95.4	94.8	-	-
	Ours	96.2	96.3	96.3	95.4	92.2 ± 2.6
	Ours + Cross-Dataset	96.3	96.0	96.1	95.2	91.9 ± 3.1

Table 4. State-of-the-Art Comparisons. We use 100% of AICrowd to train/test our final model.

network in many related works [27], and a vanilla fully-connected network since it is a notable architecture that has been tested on other building footprint datasets [22, 31]. The results indicate that the baseline method performs quite well with precision but falls behind with recall, thus creating an imbalance that brings their F1 performance down. Although the ResNet-101 backbone does not perform as

well in precision, it offers a more balanced F1 Score performance and often outperforms in mIOU. Finally, the numerical results show that the Dilated ResNet backbone outperforms the ResNet backbone across all performance metrics. The performance of the Dilated ResNet compared to the larger ResNet-101 model supports the claims laid out in Section 3.1: dilated convolutions help preserve the spa-

tial structure of footprint predictions and expand the contextual receptive field. The results of the Dilated Res-Net backbone and baseline methods can be visualized in Figure 4. Empirically, the visual results show that the FCN produces noisy footprints, especially on the border, and the U-Net often produces more "under-filled" footprints, thus the lack in recall. The Dilated Res-Net backbone (labeled "BCE+1") provides significantly sharper footprints notwithstanding the conjoined predictions.

5.3. β in F-Beta Measure and Exponentially Weighted Boundary Loss

Figure 3 illustrates how precision, recall, and overall F-1 score vary as we tune the β in the F-Beta measure. Reinforcing the claims in Section 3.2, the figure shows that recall increases with β , but at a much slower rate compared to how precision proliferates as β decreases. From this figure, we claim that building segmentation is inherently a problem of precision, and overall performance benefits from placing more emphasis on precision than recall. We note that in addition to precision's faster rate of growth, as we bring β closer to zero, the overall F-1 score increases exponentially.

We report the results of using the F-Beta Measure and the Exponentially Weighted Boundary Loss (EWC) in the form of two tables. First, Table 2 concerns the ablation study between F-Beta and binary cross entropy (BCE). The data shows that the network does not benefit from only BCE, and the performance of using only the DICE measure ($\beta = 1$) falls slightly short of the performance reached by both losses combined. Furthermore, the table illustrates the trade-off between low and high values of β when combining the BCE and F-Beta objective functions. With high β , we achieve much better recall but suffer in all other metrics. With $\beta = 0.1$, we achieve considerable gains in precision and mIOU, and despite the lack in recall we consistently achieve a higher F-1 score.

In Table 3, we report the results of F-Beta and EWC. We first note that EWC alone performs slightly worse than BCE, despite placing more emphasis on separating buildings. In addition, EWC tends to achieve higher performance in recall than precision. This is because while EWC is focused on boundaries, it is not focusing on preventing sparse, erroneous footprint predictions. On the other hand, we incur overall performance benefits when we combine EWC and F-Beta together. The relative precision and recall trade-offs of β are still maintained when using EWC, and $\beta = 0.1$ continues to have better overall F-1 and mIOU performance. Besides the increase in mIOU, when using EWC over BCE in conjunction with F-Beta we generally observe that precision tends to fall and recall improves. Consequently, this creates a more balanced performance between precision and recall, thus consistently improving the F-1 score. Figure 4 shows the results of using EWC and $\beta = 0.1$. Com-

pared to the previous iteration, we effectively resolved previously conjoined and sparse false-positive footprints across all three datasets. For the footprints that are still conjoined, we have improved separation.

5.4. Cross-Dataset Training

Table 4 reports the performance of our final network when using cross-dataset training. Note that despite inter-dataset variance, as illustrated by Figure 1, the cross-dataset strategy achieved a slightly higher performance on both Urban3D and SpaceNet. The same strategy did not perform similarly on the AICrowd dataset, which leads us to believe that cross-dataset training is not as effective on datasets with a saturated presence in the combined set. Qualitatively, Figure 4 shows that our final method generates footprints with a more defined shape and fewer conjoined predictions.

6. Conclusion

In order to better support urban energy simulation through Deep Learning based approaches, convolutional neural networks must first be able to produce quality results with distinct footprints, and be scale invariant to the fickle nature of satellite imagery. To increase accessibility to this technology, approaches should also be robust to the lack of special wideband imagery that only few urban centers are able to produce.

In this paper, we present a new approach that brings Deep Learning closer to perfecting building segmentation. In the presence of class imbalance, we demonstrate that our proposed F-Beta measure is efficacious at resolving false-positives without de-prioritizing the common class. We show that combining F-Beta with our EWC loss creates a powerful tool to separate distinct entities that were previously unified. By using these advances and a DeeplabV3+ architecture with a Dilated ResNet backbone, we achieve state-of-the-art performance on three standard datasets and provide a strong network which can assist urban planners worldwide in climate related decision making.

References

- [1] Myles Allen, Mustafa Babiker, Yang Chen, and Heleen C. de Coninck. *IPCC SR15: Summary for Policymakers*. Intergovernmental Panel on Climate Change, Oct. 2018. [1](#)
- [2] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Joint learning from earth observation and openstreetmap data to get faster better semantic maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. [1](#)
- [3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. [2](#)
- [4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous

- separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 3, 4, 6
- [5] Sean Andrew Chen, Andrew Escay, Christopher Haberland, Tessa Schneider, Valentina Staneva, and Youngjun Choe. Benchmark dataset for automatic damaged building detection from post-hurricane remotely sensed imagery. *arXiv preprint arXiv:1812.05581*, 2018. 1
- [6] Jakub Czakon, Kamil Kaczmarek, Andrzej Pyskir, and Tarasiewicz Piotr. Mapping challenge winning solution, September 2018. [Online; posted 12-Sept-2018]. 6, 7
- [7] Remi Delassus and Romain Giot. Cnns fusion for building detection in aerial images for the building detection challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 242–246, 2018. 1, 2, 3, 6, 7
- [8] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 2, 3, 4, 5
- [9] Chao Ding and Khee Poh Lam. Data-driven model for cross ventilation potential in high-density cities based on coupled cfd simulation and machine learning. *Building and Environment*, 165:106394, 2019. 1
- [10] Chao Ding and Nan Zhou. Using residential and office building archetypes for energy efficiency building solutions in an urban scale: A china case study. *Energies*, 13(12), 2020. 1
- [11] GDAL/OGR contributors. *GDAL/OGR Geospatial Data Abstraction software Library*. Open Source Geospatial Foundation, 2021. 5
- [12] Hirsh Goldberg, Myron Brown, and Sean Wang. A benchmark for building footprint classification using orthorectified rgb imagery and digital surface models from commercial satellites. In *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7, 2017. 1, 5, 10
- [13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2
- [14] Timothy L Haithcoat, Wenbo Song, and James D Hippel. Building footprint extraction and 3-d reconstruction from lidar data. In *IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas (Cat. No. 01EX482)*, pages 74–78. IEEE, 2001. 2
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [16] Paul Jaccard. The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50, 1912. 5
- [17] Shruti Jadon. A survey of loss functions for semantic segmentation. *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Oct 2020. 3
- [18] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *International conference on medical imaging with deep learning*, pages 285–296. PMLR, 2019. 3
- [19] Weijia Li, Conghui He, Jiarui Fang, Juepeng Zheng, Hao-huan Fu, and Le Yu. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source gis data. *Remote Sensing*, 11(4), 2019. 1, 2, 3
- [20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3
- [21] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. cite arxiv:1405.0312Comment: 1) updated annotation pipeline description and figures; 2) added new section describing datasets splits; 3) updated author list. 5
- [22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 7, 10
- [23] Dimitris Marmaris, F. Adam, Mihai Datcu, Thomas Esch, and Uwe Stilla. Deep neural networks for above-ground detection in very high spatial resolution digital elevation models. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W4:103–110, 03 2015. 2, 4
- [24] Sharada Prasanna Mohanty. Crowdai mapping challenge 2018:baseline with mask rcnn. <https://github.com/crowdai/crowdai-mapping-challenge-mask-rcnn>, 2018. 1, 5, 10
- [25] Intergovernmental Panel on Climate Change. *Human Settlements, Infrastructure, and Spatial Planning*, page 923–1000. Cambridge University Press, 2015. 1
- [26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Razis, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6
- [27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4, 7, 10
- [28] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387. Springer, 2017. 3
- [29] Thorvald A. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biol. Skr.*, 5:1–34, 1948. 2, 3, 5
- [30] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow.

- Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 1, 5, 10
- [31] Hsiuhan Lexie Yang, Jiangye Yuan, Dalton Lunga, Melanie Laverdiere, Amy Rose, and Budhendra Bhaduri. Building extraction at scale using convolutional neural network: Mapping of the united states. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(8):2600–2614, 2018. 2, 7
- [32] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 3, 6
- [33] Jing Zhang, Wanqing Li, and Philip Ogunbona. Transfer learning for cross-dataset recognition: a survey. *arXiv preprint arXiv:1705.04396*, 2017. 5
- [34] Kang Zhao, Muhammad Kamran, and Gunho Sohn. Boundary regularized building footprint extraction from satellite images using deep neural network. *arXiv preprint arXiv:2006.13176*, 2020. 1, 6, 7
- [35] Stefano Zorzi, Ksenia Bittner, and Freidrich Fraundorfer. Machine-learned regularization and polygonization of building segmentation masks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3098–3105, Los Alamitos, CA, USA, jan 2021. IEEE Computer Society. 2, 6, 7

A. Supplemental Materials

A.1. Additional Visualizations

Besides Figure 4, we provide additional visualizations of model predictions on the Urban3D [12], SpaceNet [30], and AICrowd [24] datasets in Figure 5.

A.2. Further Analysis

As discussed in Section 4 of the main paper and supported by Figures 4 and 5, the predictions by the fully-convolutional network [22] frequently exhibit noisy and sparse predictions. Although the U-Net [27] resolves almost all of these noisy predictions, thus the high average precision, the model does not effectively capture the geometry of footprints which brings down the average recall. This is especially evident in Figure 5 (b)(i), where the U-Net is unable to capture the smaller buildings in the upper portion of the center neighborhood in the satellite image.

The first iteration of our network (column "BCE + 1") solves the issues the U-Net gave rise to as it is now able to capture more of the building footprints. However, as discussed in the main paper, this iteration often creates conjoined footprint predictions and curved building geometries. The next iteration of our network (column "ECW + 0.5") employs the F-Beta measure and the exponentially weighted boundary loss to encourage the network to separate distinct footprints and create sharper edges. These claims are observable primarily in the predictions of Urban3D and SpaceNet examples (Figure 5 (a) and (b)). Notice that by using a new objective, we were able to separate most of the combined footprints from the previous iteration and create straighter edges. Finally, the last iteration of our network (column "Final Network") involved using a cross-task training strategy on only the Urban3D and SpaceNet datasets to resolve any remaining prediction artifacts. Observe from Figure 5 that this strategy consistently produces predictions that improves upon the previous iteration by separating footprints that were still combined, enforces sharper corners, and cleans up most remaining false positives.

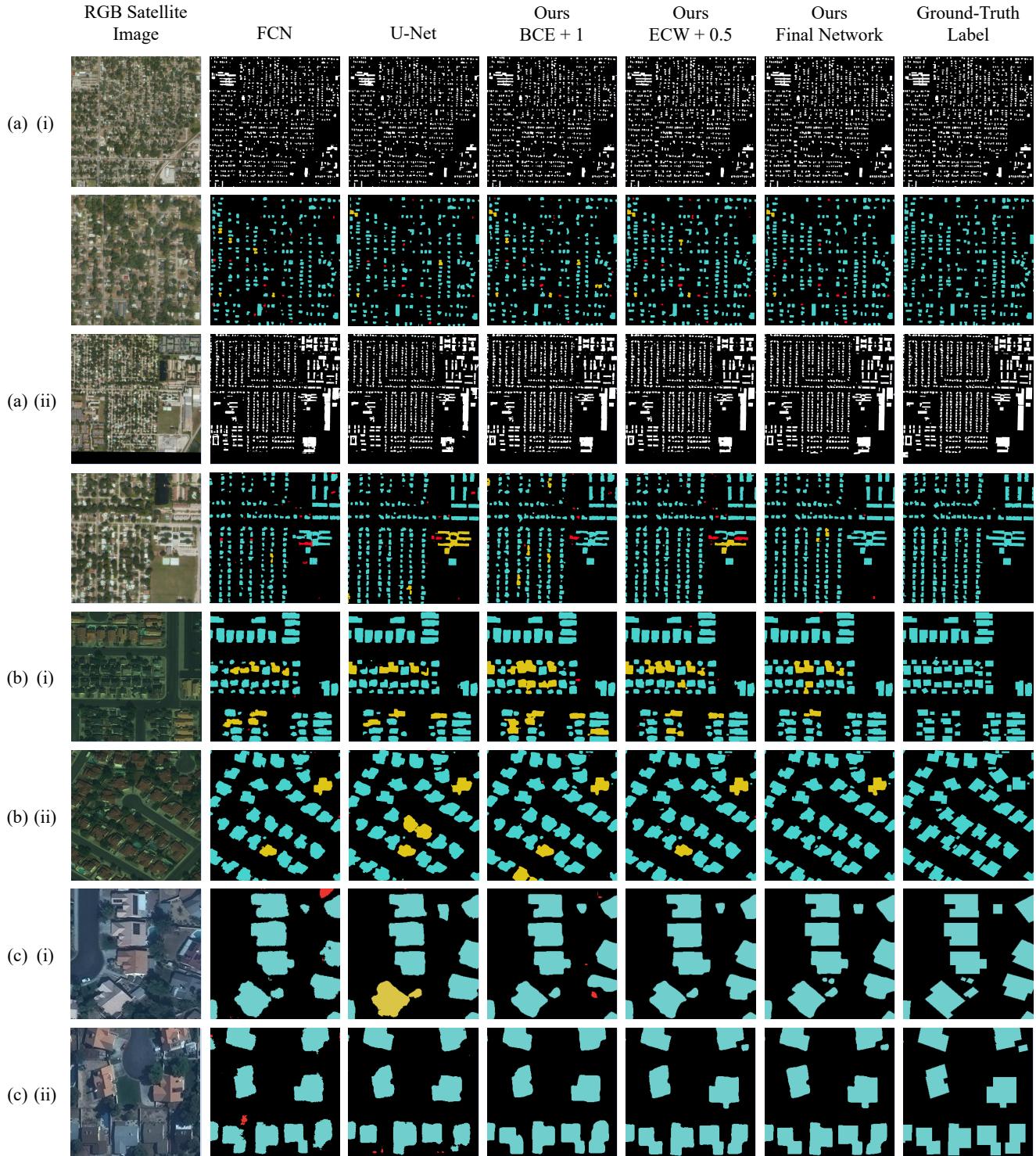


Figure 5. Additional **two** visualizations of predictions from each of the three datasets used in this paper. **Top-Down:** (a) Urban3D [the second row of each prediction is a magnification of the first row for better visual comparison], (b) SpaceNet, (c) AICrowd. Turquoise buildings are correctly predicted footprints, yellow predictions are two conjoined ground-truth footprints, and red predictions are false positive footprints.